Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

FairPlay: Detecting and Deterring Online Customer Misbehavior

Ji Wu, Zhiqiang (Eric) Zheng, J. Leon Zhao

To cite this article:
Ji Wu, Zhiqiang (Eric) Zheng, J. Leon Zhao (2021) FairPlay: Detecting and Deterring Online Customer Misbehavior. Information Systems Research 32(4):1323-1346. https://doi.org/10.1287/isre.2021.1035

http://pubsonline.informs.org/journal/isre

INFORMATION SYSTEMS RESEARCH
Vol. 32, No. 4, December 2021, pp. 1323–1346
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

FairPlay: Detecting and Deterring Online Customer Misbehavior

Ji Wu,[a] Zhiqiang (Eric) Zheng,[b] J. Leon Zhao[c]

[a]School of Business, Sun Yat-sen University, Guangzhou 510275, China; [b]Jindal School of Management, University of Texas at Dallas, Richardson, Texas 75080; [c]School of Management and Economics, Chinese University of Hong Kong, Shenzhen 518172, China

Contact: wuji3@mail.sysu.edu.cn, https://orcid.org/0000-0002-3417-635X (JW); ericz@utdallas.edu, https://orcid.org/0000-0001-8483-8713 (Z(E)Z); leonzhao@cuhk.edu.cn, https://orcid.org/0000-0002-0624-0254 (JLZ)

Received: July 16, 2019
Revised: August 5, 2020; January 30, 2021
Accepted: April 5, 2021
Published Online in Articles in Advance: September 17, 2021

https://doi.org/10.1287/isre.2021.1035

Abstract. Customer misbehavior is a serious and pervasive problem in firm-sponsored social media, yet prior studies provide limited insight into how firms should detect and manage it. To address this gap, we first develop a data science approach to detect customer misbehavior on social media and then devise intervention strategies to deter it. Specifically, we build on natural language processing and deep learning techniques to automatically detect customer misbehavior by mining customers' social media activities in collaboration with a leading apparel firm. The results show that our algorithmic solution achieves superior performance, improving detection by 7%–9% compared with traditional methods. We then implement two types of intervention policies based on the focus theory of normative conduct that advocates the use of injunctive norms (i.e., a punishment policy) and descriptive norms (i.e., a common identity policy) to restrain customer misbehavior. We conduct field experiments with the firm to validate these policies. The experimental results indicate that punishment considerably reduces customer misbehavior in the short term, but this effect decays over time, whereas common identity has a smaller but more persistent effect on misbehavior reduction. In addition, punishing dysfunctional customers decreases their purchase frequency, whereas imposing a common identity increases it. Interestingly, our results show that combining the two policies effectively alleviates the detrimental effect of punishment, especially in the long run. We examine the heterogeneous treatment effect on novice and experienced customers. Finally, a follow-up field experiment reveals that the disclosure of the use of an artificial intelligence detector improves the effectiveness of the intervention strategies, and this effect is more pronounced for the punishment and combination strategies.

History: Wonseok Oh, Senior Editor; Yili (Kevin) Hong, Associate Editor.
Funding: This research is partially supported by the National Natural Science Foundation of China [Grants 72071218, 71601190, and 71831006]. J. Leon Zhao's research is partially supported by the General Research Fund [Grant CityU 11508517] from the Research Grants Council of Hong Kong.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/isre.2021.1035.

Keywords: customer misbehavior • norm enforcement • deep learning • field experiment • punishment • common identity

1. Introduction

In June 2020, the Anti-Defamation League led a "StopHateforProfit" campaign against Facebook, citing its "repeated failure to meaningfully address the vast proliferation of hate speech on its platform." The campaign called on big corporations to pause their advertising on Facebook.[1] Within days, a number of companies, including Unilever, HP, Coca-Cola, and Starbucks, pulled ads out of Facebook in support of the campaign. Although Facebook defended the measures it has taken to fight against hate speech, including continuously working with experts to update its policies and investing billions of dollars in technology to detect policy violations, the number of companies boycotting Facebook continues to mount.

This story reveals the surprising fact that even Facebook, arguably the most successful social media firm in the world, is still grappling with detecting and managing user misbehavior such as hate speech on its platform. This brings up a fundamental question: how can firms detect and manage user misbehavior on social media? Failing to address it could lead to dire consequences, such as those confronting Facebook, considering that more than six million companies have set up brand communities on Facebook for marketing purposes (Bapna et al. 2019).

Facebook certainly is not alone. Firms are increasingly utilizing new information technologies (IT), such as social media, to interact with customers. But, in the meantime, a customer could wreak havoc on a firm when using such technologies inappropriately. Inappropriate use of IT by customers is broadly referred to as customer misbehavior. This study focuses on customer misbehavior in the use of a firm's social

media for its online brand community (OBC), in which customers interact with the firm and other consumers regarding the firm's products or services. According to Manchanda et al. (2015), almost half of the top 100 global brands host their own OBCs. We define *customer misbehavior* in our context as a customer's posting behavior that violates the rules/norms set forth by the firm for its social media, threatening the well-being of the firm and other customers. Our definition based on rule violation echoes Twitter's approach in managing posting misbehavior when it permanently suspended the @realDonaldTrump account on January 8, 2021. In justifying its action, Twitter stated that "violations of the *Twitter Rules* would result in this very course of action … These accounts are not above our rules entirely and cannot use Twitter to incite violence … "[2]

In recent years, online customer misbehavior is on the rise, and it has cost firms a large sum (Garnefeld et al. 2019). However, as the Facebook case demonstrates, although online customer misbehavior is pervasive and poses a serious threat to firms, systematic detection methods and intervention strategies to effectively manage such misbehavior are lacking.

This predicament has recently spurred great interest among academics in the psychology, organization science, computer science, marketing, and information systems disciplines (e.g., Venkatraman et al. 2018, Fombelle et al. 2020). We provide a comprehensive summary of this stream of literature in Table A1 of Online Appendix A. The literature, however, mostly focuses on specific variants of customer misbehavior, such as shoplifting and insurance fraud, and explores the drivers and consequences of such deviant behaviors, thus providing less than ideal solutions for detecting customer misbehavior and limited insight into how to intervene to effectively manage it. Only a handful of studies propose intervention strategies to prevent customer misbehavior for specific forms of customer deviance, such as wardrobing (Shang et al. 2017), typically at a conceptual level (Bhati and Pearce 2016, Fombelle et al. 2020). Empirical evidence and field experiments on the effectiveness of more general intervention strategies are lacking.

Motivated by these research gaps, we developed a closed-loop solution to (1) detect customer misbehavior using machine learning (ML) and (2) manage customer behavior using norm enforcement strategies. We demonstrate the effectiveness of our solution using field experiments. Our misbehavior detection method devises a novel overarching architecture that leverages a comprehensive set of features capturing customer deviance, enabling detection of misbehavior automatically. Our approach goes beyond the related literature on misbehavior detection, which we comprehensively review and summarize in Online Table B1. A key distinction is that our method attempts to provide a full understanding of how various forms of deviant behaviors can be detected by drawing on 360° touchpoints a customer has with the firm, including customers' historical posting activities, textual posts, demographics, etc., whereas extant studies (e.g., Pitsilis et al. 2018, Rosa et al. 2019) mostly focus on a specific aspect of customer activity (e.g., textual posts) to detect a specific form of misbehavior based on which one cannot see the forest for the trees.

We then designed intervention strategies using two types of norms—injunctive (i.e., punishment of individuals) and descriptive (i.e., establishment of a common identity within a group)[3]—to mitigate customer misbehavior. Contextually, this study extends the literature to a new OBC context in which the offenders are clients of the sufferers in a client–supplier relationship, and therefore, extra care is required because interventions could backfire and damage the relationship. Accordingly, we studied the nuanced efficacy of these intervention strategies in terms of reducing repeat violations and increasing purchases. Therein we identified the subtle trade-off between these objectives as well as the conditions under which such a trade-off occurs, an important aspect of effective management of customer misbehavior that has not been explored in the literature.

Methodologically, we combined ML (to detect customer misbehavior) with field experiments (to study the efficacy of the intervention strategies). We first collected archived customer behavior data from the OBC of a leading online clothing firm and assembled a unique data set combining customers' posting activities and personal profiles. Then, we developed an algorithmic approach, *FairPlay*, to identify violating posts in the OBC using natural language processing (NLP) and deep learning (DL) techniques. In the field experiment, we used the algorithm to monitor and flag customer misbehavior. Through the OBC, we then sent each misbehaving customer a private message, chosen randomly from one of three types of normative messages: a penalty notification, the common identity that members hold in the community, or a combination of the two. A control group of randomly chosen customers received an acknowledging message without any intervention. We tracked customer behavior in the months after the release of messages and assessed the effectiveness of the interventions in terms of two outcomes: (1) subsequent violations to show how effective the messages were in deterring a customer's future misbehavior and (2) purchase frequency to assess how the messages affected a customer's future purchases. We also designed a follow-up field experiment to explore how normative messages with the disclosure of the use of an artificial intelligence (AI)/ML[4] detector affected the two outcomes of misbehaving customers.

The results show that the proposed algorithmic solution is effective in detecting online customer misbehavior with a satisfactory macroaverage F_1-score in the range of 82% to 86%, outperforming various benchmark methods. Among the algorithms we considered, a convolutional neural network built on all of the feature categories yielded the best performance. Moreover, the results from our field experiments suggest that injunctive and descriptive norms are effective in different ways. Specifically, punishment reduces customer misbehavior in the short term, but its effect decays over time and dissipates in the long term. In contrast, we observed a relatively smaller but more persistent reduction in violations among customers who received messages regarding the common identity of the community. Our results also reveal that the combined policy of punishment and common identity achieves larger and longer-term reductions in customer misbehavior. Our results also show that punishing dysfunctional customers can decrease their future purchases, whereas enforcement via common identity can significantly increase their future purchases. Moreover, we find a heterogeneous treatment effect in that punishment and the combined approach are more effective with novice customers, whereas common identity works better with experienced customers. Finally, the results show that the disclosure of the use of an AI detector improves the effectiveness of normative messages, and this positive effect is more pronounced for the injunctive and combined norm approaches than the descriptive norm approach.

2. Literature Review

2.1. Defining Customer Misbehavior

The literature has used various terms to refer to customer misbehavior, such as "deviant customer behavior" (Bhati and Pearce 2016), "problem customer behavior" (Bitner et al. 1994), "unethical customer behavior" (Mitchell et al. 2009), and "dysfunctional customer behavior" (Yang et al. 2017). Online Appendix A presents an extensive overview of these diverse perspectives.

Prior studies primarily focus on examining a specific form of customer misbehavior and define it as a particular deviant action (e.g., Shang et al. 2017). Fombelle et al. (2020) argue that treating different forms of deviant actions as different phenomena results in a lack of collective understanding of customer misbehavior. In this regard, Fullerton and Punj (2004) propose a systematic approach to capturing customer misbehavior from the perspective of norm violation: first defining the norms (with which customers should comply) and then identifying customer behaviors deviating from those norms as the misbehaving ones. This approach is advocated by Daunt and Harris (2012), who assert that the "norm-based approach is the most appropriate mechanism for conceptualizing and measuring incidents of dysfunctional customer behavior" (p. 131). We follow this preach in this study and define customer misbehavior as a customer's posting behavior in a firm's social media that violates the norms stipulated by the firm. We focus on posting behavior because studies show that such behaviors determine the success or failure of online community and have become the major concern for managers on social media management (Bapna et al. 2019, Chung et al. 2020). This definition specifies the scope of our misbehavior investigation, which emphasizes a customer's deviant IT use (posting activity in the firm's social media) and norm infringement. This answers the call from Venkatraman et al. (2018) for a systematic investigation of deviant behavior from the perspective of users' deviant IT use.

2.2. Gaps in the Misbehavior Literature

Prior studies have sought to provide insights into general personality-based motivations (e.g., Daunt and Harris 2011) and contextual triggers (Lowry et al. 2016) for customer misbehavior and have examined the consequences of customer deviance, such as the negative influence on a firm's performance and service process (Warren and Schweitzer 2018), deleterious effects on frontline employees (Wang et al. 2011), and deterioration of the other customers' satisfaction (Bitner et al. 1994). However, these studies stop short of answering how firms can detect and deter customer misbehavior.

In some industries, such as the insurance industry, firms rely on manual audits and preimposed heuristics in identifying customer misbehavior in an off-line setting. For instance, Warren and Schweitzer (2018) find that claimant interviews are the most important vehicle in discerning insurance claim fraud using prespecified rules of thumb. Marketing scholars investigating the dark side of customer behavior question the effectiveness of such measures (Garnefeld et al. 2019) and conclude that detecting customer misbehavior remains a critical managerial problem that requires automatic detection methods. Recently, studies in computer science have attempted to measure and detect specific types of bad behaviors, such as cyberbullying (Tahmasbi and Rastegari 2018), hate speech (Djuric et al. 2015), spam (Pitsilis et al. 2018), and harassment (Yin et al. 2009), on platforms such as Yahoo (Nobata et al. 2016), Twitter (Rosa et al. 2019), and YouTube (Chen et al. 2012). A summary of the approaches used in misbehavior detection, along with their characteristics and performance, is presented in Online Appendix B.

Online Table B1 shows that most previous studies have relied on traditional machine learning classifiers,

such as logistic regression (Djuric et al. 2015), support vector machines (Rosa et al. 2019), and random forests (Chatzakou et al. 2017), or ensemble classifiers of such traditional methods (Pitsilis et al. 2018) to detect customer misbehavior. Some recent studies have started experimenting with deep learning and found it promising.

Online Table B1 also lists the input features used in each study. Four types of features are commonly used: text, content, context, and user. Among text-based features, *N*-grams, term frequency, and word embeddings are the most common. Content-based features typically contain information extracted from the contents of posts (e.g., sentiment, topic, etc.), whereas context-based features are related to a user's behavioral pattern (e.g., posting frequency). Finally, user-based features, such as age and gender, are extracted from a user's profile. Text-based features are the predominant category used to capture misbehavior (Djuric et al. 2015, Nobata et al. 2016). However, prior studies point out that text-only features alone may be inadequate to reliably detect misbehavior (Yin et al. 2009, Chen et al. 2012). Nevertheless, very few studies combine text-based features with other types of features when identifying deviant behavior (Pitsilis et al. 2018, Singh et al. 2018). Surprisingly, no study has leveraged all four feature categories, perhaps because of the scope of the literature or data limitations. Moreover, studies mostly focus on detecting a specific form of online misbehavior (e.g., cyberbullying or harassment). This study tackles the rather complex misbehavior detection problem more generally, which encompasses a broad array of expressions and characteristics in OBCs. Using all of the feature categories helped us capture different facets of user behavior and, thus, enabled us to detect a wide array of misbehaviors.

Although it is widely recognized that customer misbehavior can lead to dire consequences for businesses, there is a lack of empirical research or practical guidance on how firms should deter customer misbehavior. A summary of the related studies and their proposed intervention strategies are presented in Online Table A1. Some studies examine specific intervention strategies for a specific form of customer misbehavior, such as insurance fraud or wardrobing (Shang et al. 2017, Garnefeld et al. 2019), and thus, their intervention strategies are relatively context contingent. For example, Shang et al. (2017) investigate how retailers can set prices and refund policies to prevent wardrobing. A few studies have addressed the managerial use of education, deterrence, or environmental design to prevent deviant behavior. However, such studies typically are either restricted by their reliance on qualitative data or focused on an off-line setting. For instance, Fullerton and Punj (2004) propose using education and deterrence to contain customer misbehavior. However, their study is descriptive in nature and only considers education and deterrence at the macrolevel without specifying how the constructs are to be operationalized. Nor does it offer empirical evidence on how such approaches could affect dysfunctional customers. In contrast, our study specifies the prevention strategies of a firm, devises optimal interventions for heterogeneous customers, and investigates their effectiveness through experiments. Another closely related work is Yang et al. (2017), which studies customer misbehavior using observational data to investigate how tolerating unethical customers can adversely influence a retailer's long-term revenue. Their work only considers toleration as a response strategy. We extend their research by directly designing intervention strategies that help correct customer misbehavior via norm enforcement.

2.3. Norm Enforcement Mechanism (Policy)

We take the norm-based approach to defining customer misbehavior. By our definition, customer misbehavior is the infringement of norms. Therefore, the goal of misbehavior management is to enforce norms to discourage customers from deviating from them. The *focus theory of normative conduct* suggests that there are two approaches to establishing and reinforcing norms: (1) descriptive norms can be enforced through establishing a common identity within a group, and (2) injunctive norms can be enforced through punishment (Christensen et al. 2004, Cialdini et al. 2006, Weng and Carlsson 2015). Although the literature documents that common identity and punishment can be used to prevent unethical behaviors, these behavioral mechanisms remain untested in our new context, in which violations occur in an OBC and the offenders are clients of the sufferers in a client–supplier relationship.

Injunctive norms refer to how individuals ought to behave to avoid penalties (Cialdini et al. 2006). Punishment is shown to be a necessary and important instrument to reduce undesirable behavior (Nagin 2013). The economic theory of deterrence, which has been widely applied in the fields of criminal justice, economics, and law, suggests the prospect of punishment in deterring bad behavior (Fehr and Gachter 2000). However, there is no consensus regarding the consequence of punishment on customer–firm relationships. Some studies suggest that organizations ought to use punishment to discourage undesirable behavior of customers (Kim and Smith 2005). Others, however, argue that punishing customers backfires, leading to reprisals and increased customer violations (Balafoutas et al. 2014, Golf-Papez and Veer 2017). Moreover, Kim and Smith (2005) find, in a survey of customer responses to service organizations' penalties, that imposing a penalty could lead customers to

devalue a service organization's warmth and decrease their loyalty. In contrast, Tax and Nair (2013) suggest that firms could benefit from punishing misbehaving customers without harming customer value. Overall, the effectiveness of punishment in regulating customer misbehavior is inconclusive and even controversial, and understanding it requires systematic empirical investigations. We conduct such an investigation by estimating the effects of punishment on dysfunctional customers regarding their future violations and purchases. In addition, we investigated how such effects differ in the short and long term and for heterogeneous dysfunctional customers.

Descriptive norms refer to what most people typically do to acquire identity relevance (Christensen et al. 2004). Identity, a person's sense of self, is one of the key driving forces behind individual behavior (Akerlof and Kranton 2000). A person's sense of self is embedded in the person's social status (class) and is influenced by the collective behavior of people in that class. *Social identity theory* (Tajfel 2010) suggests that, once an individual has gone through a process to categorize the self as part of a unit with shared goals, values, and norms, the individual's behavior tends to conform to the norms of the unit, which leads to greater compliance (Akerlof and Kranton 2000). A number of studies in experimental economics show that salient identification can reduce unethical behavior (i.e., free riding) in the public good setting (McLeish and Oxoby 2011). Algesheimer et al. (2005) find that, in brand communities, customer identification with an organization tends to result in assimilation of the community's norms, values, and goals. Raïes et al. (2015) show that customer identification with an organization improves customer commitment and, subsequently, customer value. However, to the best of our knowledge, the enforcement mechanism of identity remains untested in real-world business practice, in which firms use it to counter client misdeeds.

Both punishment and identity are considered instrumental, and they are studied in off-line (physical) contexts. However, researchers acknowledge that theories and studies of traditional violations are not applicable to cyberdeviance because of the differences between online and off-line contexts (Lowry et al. 2016). Thus, an unexplored issue is to what extent these intervention strategies can be used to restrain OBC violations. In addition, Cialdini et al. (2006, p. 13) suggest that "norm-based persuasive communications are likely to have their best effects when communicators align descriptive and injunctive normative messages to work in tandem." However, to the best of our knowledge, no field experiments have explicitly examined the combination of punishment and common identity in affecting prosocial behavior. In this study, we designed field experiments and empirically examined the effects of different normative messages (i.e., punishment, identity, and their combination) on deterring online customer misbehavior toward achieving better economic outcomes, such as fewer future violations or more purchases.

2.4. Challenges in Studying Misbehavior

Studying customer misbehavior presents serious challenges because of the difficulty of data collection, processing, and analysis (Fombelle et al. 2020). Research on customer misbehavior typically relies on self-reported survey data or limited data from small-scale laboratory experiments under hypothetical or even unethical conditions. Such data collection schemes are questionable. For instance, in the survey approach, to gauge clandestine consumer activities, researchers may need to ask respondents to recall or imagine socially undesirable or unethical behaviors (Daunt and Harris 2011). Such a questionnaire may cause offense, embarrassment, or stress to subjects (Brinkmann and Lentz 2006), leading to social desirability bias and compromised data quality.

Our study overcame these challenges by resorting to the use of a firm's objective archival data in conjunction with field experiments within that firm's OBC. Specifically, we assembled a unique data set that tracks customer activities on the firm's social media platform, customer profiles, and their historical transactions. Using this unique data set, we were able to identify real-life customer misbehavior using ML techniques.[5] We then operationalized punishment- and identity-based normative messages for the firm to counter customer misbehavior. Our randomized field experiment spearheaded an ideal means to address potential endogeneity in isolating the effects of these intervention strategies.

3. Research Data and Context

We collaborated with a leading apparel firm in China to conduct the field experiment.[6] The firm's products are private-label products that are only available through its retail channels. The focal firm operates in an omni-channel mode, primarily through its online channel plus a small off-line presence (with a small number of physical stores). The firm carries eight product categories, including tops, shirts, and coats. The firm is the largest retailer by sales volume in its core product category and has an annual revenue of more than US$200 million.

The firm launched an OBC in November 2013 to connect with and engage its customers to increase customer awareness and loyalty to the firm's brand. The OBC allows customers to create a personal profile and engage in the community by posting, commenting, and communicating with others. In addition,

registered customers can log into both the e-commerce platform and the specific brand community using the same account, which allows tracking of customers' transaction records.

The customer community provides an ideal test bed for our experiment. First, the OBC is fraught with customer misbehavior, including poaching campaigns about competing brands, verbal abuse of other customers, and public attacks on the brand. Second, the brand community has the will and full authority to implement our norm enforcement strategies, for example, punishing offenders by forbidding and deleting illegal contents. Customers engaged in the OBC tend to build a common identity with the community or the brand (Algesheimer et al. 2005, Ren et al. 2012), which allowed us to readily test the identity-based mechanism.

Figure 1 summarizes our data collection process and the experimental design. On June 1, 2016, there were 14,059 customers in the OBC. We used the pool of community participants who had registered before June 2016, who were active (posted at least once) between June 2016 and November 2016, and who had an active email address so that the firm could reach them. The June 2016 cutoff ensured at least six months of preintervention data for all the customers in the study. Among all the community participants, 2,435 met our screening criteria (see Figure 1).

4. Customer Misbehavior Detection

4.1. FairPlay: A Framework for Misbehavior Detection

We propose an integrative analytics framework, FairPlay, to automatically identify customer misbehavior in the firm's OBC. Departing from previous studies that have used either textual or custom features in identifying individual forms of misbehavior, we focused on a rather general misbehavior detection problem in which misbehavior could be exhibited in a wide array of forms. We designed a novel overarching architecture that leveraged all of the available features that collectively paint a rather complete picture of customer misbehavior, to detect a broad set of customer misbehaviors. Figure 2 presents FairPlay and depicts how we used the mixture of diverse data (e.g., sequences of text and nonsequential metadata). The architecture consists of two paths: the text and metadata paths. The text path extracted text-based features, whereas the metadata path extracted content-, context-, and user-based features.

The text path only considered raw text as input. Early research used bag-of-words (BoW) to construct text-based features in cyberbullying or spam detection (Tahmasbi and Rastegari 2018). The BoW approach treats text as an unordered collection of words, disregarding semantic information. Word embedding is an extended version of BoW that considers contextual information about words. To account for product-category and source-specific words, we created word embedding from our corpus of OBC posts. We used *fastText* that jointly models posts and words, with which we learn their representation distribution in a joint space using the skip-gram model. The skip-gram model is a predictive model that maximizes the average likelihood of words appearing together in a sequence of l words:

$$\frac{1}{I}\sum_{i=1}^{I}\sum_{-l\le j\le l, j\neq 0}\log p(w_{i+j}\mid w_i), \tag{1}$$

$$p(word_j|word_i)=\frac{\exp(v_j v_i')}{\sum_{k=1}^{|V|}\exp(v_k v_i')} \tag{2}$$

where I is the number of words in the corpus, V denotes the set of all feasible words in the vocabulary, and v_i represents the d-dimensional real-vector embedding. For our application, we set the vector size of word embedding as $d = 100$ and the size of the context vector as $l = 5$.[7]

After obtaining the word embedding of a post, we then constructed its sentence embedding using a composition function, which is a mathematical process for combining multiple words into a single vector (Iyyer et al. 2015). For simplicity, we used the unweighted averaging composition and created sentence embedding by averaging the word embedding scores corresponding to the words in the sentence.[8] Word

Figure 1. Timeline of Our Study

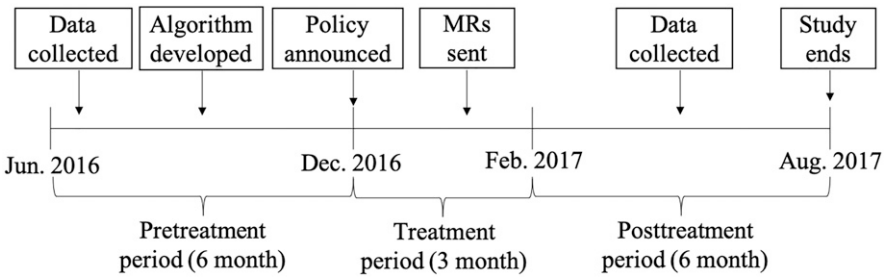
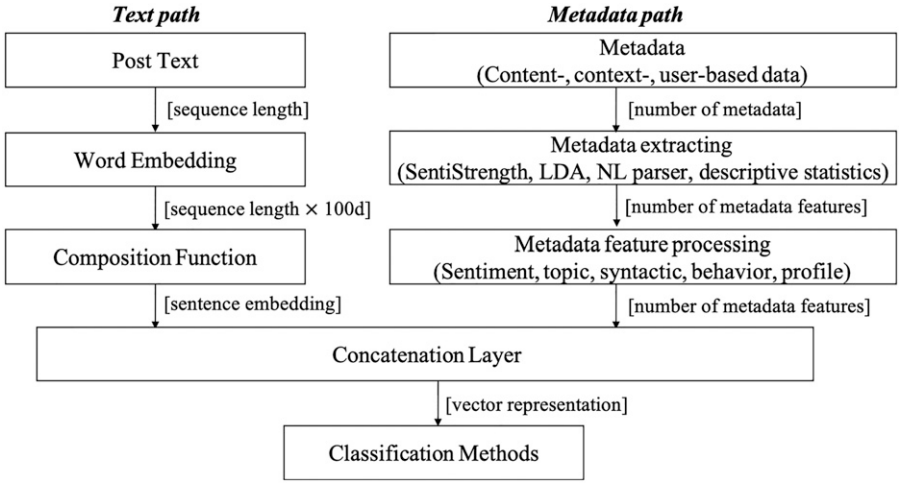


Figure 2. FairPlay: A Framework for Customer Misbehavior Detection



averaging has been advocated for its consistent and good performance (e.g., those by Wieting et al. 2016) and for its ability to produce a generic representation of posts for downstream supervised tasks (Arora et al. 2017).

The metadata path deals with nonsequential data. These metadata are of three general categories: content-, context-, and user-based. First, for content-based metadata, we considered the most common metrics describing a user's post: sentiment, topic, and syntactic structure. Prior studies use sentiment scores or post topics to distinguish between bullying and nonbullying (Nahar et al. 2013). Second, context-based metadata are highly related to a user's behavioral pattern, and studies demonstrate that such action-wise features provide important context for discerning misbehavior (Li et al. 2017). We operationalized context-based metadata as past posting frequency, post intervals, average post length, special character usage, and the number of violating posts, which collectively reflect users' engagement behavior in social media over time. Finally, for the user level, we extracted several key metadata regarding a user's demographics (i.e., age, gender, and income) and relationship with the platform (i.e., community tenure and registration type). Tahmasbi and Rastegari (2018) highlights the importance of user-based metadata in detecting misbehavior and argue that a user's profile is related to personality traits, which can be drivers of customer misbehavior (Fombelle et al. 2020). We used various methods to extract metadata features, including the *SentiStrength* algorithm, latent Dirichlet allocation, a natural language parser, and descriptive statistics. We also conducted data preprocessing, such as one-hot encoding and normalization, to transform these metadata features.

FairPlay then concatenated all of the features (see Table 1) from each path to form the final vector representation before feeding them into various classification methods. To identify which classification methods were more appropriate in our context, we first compared the state-of-the-art classification methods used by previous studies for text classification and cyberbullying detection (Djuric et al. 2015, Rosa et al. 2019). We evaluated five popular models: logistic regression (LR), support vector machines (SVM), random forest (RF), a recurrent neural network (RNN) with a long short-term memory (LSTM) cell, and a convolutional neural network (CNN) in term of their performance with respect to precision, recall, and F_1-score (F_1). In addition to these evaluation metrics, we used the overall accuracy (acc) and macro-average F_1-score ($F_{1,\ \text{macro}}$) to evaluate detection outcome. We adopted the macroaverage F_1-score because prior research shows that, in the context of violation detection, there are key advantages in using macroaveraged values from the two classes (Rosa et al. 2018).

4.2. Data Preparation

As noted, we collected the posting data of 2,435 customers between June and November 2016. There were 51,136 posts during that period. In addition, the focal firm provided us with detailed customer registration and demographic information. We assembled a unique data set that matched customers' profile data with their posts in the OBC through their user IDs. Before feeding the data into FairPlay, we conducted standard text preprocessing, including tokenization, stop word removal, and the removal of punctuation marks, numerical digits, and special symbols after tokenization. Table 1 tabulates the definitions and

Table 1. Feature Definition

Feature	Description of feature	Descriptive statistics
Text-based features (total: 100)		
Word vector	100-dimensional vector embedding	
Content-based features (total: 41)		
Ratio of nouns	Ratio of nouns in post	0.264 (0.246)
Ratio of verbs	Ratio of verbs in post	0.218 (0.207)
Ratio of pronouns	Ratio of pronouns in post	0.039 (0.083)
Ratio of adjectives	Ratio of adjectives in post	0.097 (0.185)
Ratio of adverbs	Ratio of adverbs in post	0.060 (0.102)
Function words	Number of function words in post	1.095 (1.878)
Punctuation marks	Number of punctuations in post	5.186 (6.872)
Positive sentiment	Positive sentiment score of post	3.725 (1.932)
Negative sentiment	Negative sentiment score of post	2.219 (1.437)
Overall sentiment	Overall sentiment score of post	2.604 (1.460)
Subjectivity	Subjectivity score of post	0.762 (0.556)
Topics	30 latent topics from latent Dirichlet allocation	
Context-based features (total: 11)		
Number of past posts	Total number of user's past posts	24.688 (31.956)
Maximum intervals	Maximum interarrival days of user's past posts	77.354 (97.573)
Minimum intervals	Minimum interarrival days of user's past posts	1.054 (2.682)
Average intervals	Average interarrival days of user's past posts	6.103 (10.039)
Maximum length	Maximum length of user's past posts	52.701 (93.628)
Minimum length	Minimum length of user's past posts	10.632 (17.756)
Average length	Average length of user's past posts	37.067 (46.904)
Number of emoticons	Average number of emoticons in user's past posts	2.363 (4.182)
Number of URLs	Average number of URLs in user's past posts	0.853 (1.275)
Phone number	Average number of phone numbers in posts	0.468 (0.539)
Violating posts	Number of past violating posts	1.142 (3.720)
User-based features (total: 7)		
Age	User's age	28.133 (5.642)
Gender	Whether the user is female	0.743 (0.365)
Community tenure	Number of months since user's register	20.061 (6.277)
Income	A zip code–level estimate of household income	1.564 (0.547)
South	Whether the user lives in south China	0.674 (0.430)
Off-line member	Whether the user is also an off-ine member	0.225 (0.418)
Register type	User's register device (1: PC, 2: phone, 3: tablet)	1.975 (0.279)

Note. Standard deviation in parentheses.

descriptive statistics of the features constructed by our detection architecture.

Each post in the pretreatment period was manually labeled by a group of three experts (coders) from the company. These coders were employees dedicated to managing the firm's OBC and, thus, had knowledge of the various forms of violating behaviors in the OBC. Before the coders were asked to label the posts, we provided them with a decision list (i.e., the posting guidelines in Figure 3) and ensured that they fully understood and agreed on the criteria for violating posts. After we ensured that the coders were well trained and had consistent perceptions on violating posts, each coder annotated posts independently. For our labels, we used one-hot vectors (i.e., violating or not) as the output of our system is a recommendation on whether to moderate a post rather than an empirically distributed score. We evaluated the reliability of the coders using the intercoder agreement. Our data achieved a Cohen's kappa of 0.69, suggesting a reasonably high level of consistency across the coders (Shriver et al. 2013). We used a majority vote to determine the final label. The three human coders determined that 8.03% of the posts were violating (see Figure 3).[9]

The percentage of violating instances in our data set was less than 10%. This runs in the classic problem of an imbalanced class distribution, which may negatively affect the accuracy of prediction models. We, thus, used the *synthetic minority oversampling* technique, which is appropriate when there are disproportionally few instances of positive ones (Chawla et al. 2002) to account for imbalance. We simultaneously oversampled (i.e., created synthetic instances of the minority class) and undersampled (i.e., a resampling technique without replacement) as doing so has been proven to result in better overall performance (Tahmasbi and Rastegari 2018). After randomly

Figure 3. The Released Posting Norms

Posting Guidelines on Firm X's Community

We are committed to creating an online brand community that encourages member interactions and self-expression. However, to ensure the quality of the conservations and the integrity of community members, all the members must comply with the following regulations and guidelines, and please ***do not*** post any content that:

- is defamatory, hateful, ethnically, or otherwise biased or offensive, unlawfully threatening or unlawfully harassing to any member, vulgar, indecent, obscene, or invasive of another's privacy
- promotes or fosters discrimination on the basis of race, age, religion, gender, regional origin, physical or mental disability or sextual orientation
- infringes the copyright, trademark, privacy rights, or any other legal or moral rights of any third party
- slams our brand or competitors unfoundedly
- is maliciously false or misleading
- constitutes any unsolicited or unauthorized advertising, promotional materials, junk message, spam, pyramid schemes, or any other form of solicitation

splitting the data into 90% for training and 10% for testing, we proceeded with the balancing of the training set. Online Appendix C reports the number of instances of both violating and normal posts before and after oversampling in the training set. Note that there was no resampling of the test set. Finally, we trained our classifiers with the manually labeled posts and their corresponding features.

4.3. Results of Misbehavior Detection

We systematically evaluated the efficacy of five ML techniques—RNN, CNN, LR, SVM, and RF—with tenfold cross-validation for each. The hyperparameters of the neural networks were chosen following standard approaches as proposed in previous studies (Kim 2014, Iyyer et al. 2015). We used a bidirectional RNN model with 100 hidden layers[10] for each LSTM cell. Our CNN iterated over a given list of kernel sizes and concatenated their outputs from two layers: a convolutional layer that computed features using filter windows h_t of three, four, and five with 100 feature maps each and a pooling layer with a max-pooling operation. Dropout was applied at a rate of 0.5 to prevent overfitting. These values were chosen via a grid search on our data set. The neural networks used ReLu as the activation function and softmax cross entropy as the loss function. Online Appendix D provides all the necessary details on the integrative DL architecture, its hyperparameter tuning, and the convergence of the loss function. For the SVM classifier, we adopted *SVMlight* with a linear kernel and parameterized the lowest cost parameter of SVM to optimize and balance both precision and recall. Finally, to build the RF model, we tuned the number of trees to be generated as 10 and the maximum depth as unlimited.

We report our evaluation metrics on both the misbehavior and normal class performances. Table 2 shows the label-specific precision, recall, and F_1-score values with overall accuracy and the macroaverage F_1-score. All of these values were averaged over 10 folds. The results show that the two DL methods achieved macroaverage F_1-scores in the range of 82% to 86%, outperforming traditional algorithms. These results demonstrate the superiority of a DL algorithm in detecting misbehavior. Moreover, from the experiments, we can conclude that CNN worked best for our detection task. We believe there are two underlying reasons for these findings. First, traditional methods rely on human-crafted features, and it is difficult to expect such designed features to identify diverse forms of online misbehavior. In contrast, the DL technique largely relieves the efforts on feature engineering and can learn more high-level and meaningful features automatically, enabling it to perform well in detecting generic online misbehavior. Second, the RNN architecture specializes in tackling sequential data. In our FairPlay detection architecture, we considered both sequential and nonsequential data. Effectively, our architecture decomposes patterns of customer misbehavior into a hierarchy of features with both low- (i.e., the word sequence in a post) and high-level features (e.g., posting frequency, a user's age). A CNN has the capacity to automatically extract and learn this hierarchy of features for pattern

Table 2. Tenfold Cross-Validation Result Scores in Percentage for the Misbehavior Detection Task

	Label "violation"			Label "normal"			Overall	
	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy	$F_{1,macro}$
LR	0.602	0.766	0.674	0.679	0.494	0.572	0.630	0.623
SVM	0.733	0.719	0.726	0.722	0.739	0.731	0.729	0.728
RF	0.741	0.782	0.761	0.781	0.743	0.763	0.760	0.761
CNN	0.858	0.825	0.842	0.843	0.898	0.870	0.869	0.856
RNN	0.887	0.769	0.825	0.783	0.889	0.833	0.828	0.829

recognition (Pinaya et al. 2020), explaining why CNN outperformed RNN in FairPlay.

Table 3 compares our results with the state-of-the-art baseline models as reported in recent publications. First, we evaluated our method by comparing it with benchmark models. As the baseline, we used XGBoost models with *n*-grams in detecting violating posts. XGBoost is a scalable tree-based boosting system that achieves the state-of-the-art performance in many ML challenges (Chen and Guestrin 2016). *N*-grams represent subsequences of *N* continuous words in texts, and its approach has commonly been used in information retrieval tasks, such as spam detection and sentiment identification, and has been successful for these activities (Chen et al. 2012). In our experiment, one-, two-, and three-grams were considered. In order to keep the dimensions to a reasonable scale, we filtered out grams with relatively low frequencies, leaving us with approximately 2,000 *N*-grams. We experimented with the XGBoost model in conjunction with various *n*-grams (i.e., one-, two-, and three-grams and combined grams). The results are reported in Online Appendix E. We take the XGBoost model with a unigram as the benchmark because it achieves the highest F_1-score value 75%. Table 3 shows that our approach outperforms this benchmark method. We also present the results of our experiments along with the reported results from the state-of-the-art solution proposed by Dorris et al. (2020) in Table 3. The results also demonstrate the superiority of our approach.

To fully evaluate the effectiveness of our architecture and test the incremental contribution of each category of features in a horse-race manner, we experimented with different feature classes and calculated their importance. The area under the curve (AUC) results are presented in Table 4. We observed that models built on individual feature sets performed the worst, ranging from 0.54 for user- to 0.78 for text-based features. However, performance improved when two or more types of features were combined. This indicates that each feature class adds unique information to enhance misbehavior detection. Among the three categories of metadata features, Table 4 shows that combining text- with context-based features achieves higher performance, indicating that it is necessary to include both the speech- and action-wise behaviors, echoing Li et al. (2017). A model built on all of the features yielded the best performance statistically. The results demonstrate the effectiveness of the proposed method.

To ensure that the performance of our architecture is robust to alternative data processing and analytics approaches, we conducted several robustness checks. We utilized different approaches to construct embedding features for posts and extracting metadata features. First, instead of using word averaging composition, we computed the weighted average of the word embedding in a sentence to create sentence embedding (for details, see Online Appendix F). The results reported in Online Appendix F indicate that using weighted word averaging composition does not improve detection performance. Furthermore, for the DL methods, we also combine word embedding with metadata (with or without using the bottleneck technique) in detecting customer misbehavior. The results reported in Online Table F1 show that our integrative approach performs well compared with other approaches.

Finally, our algorithms may be vulnerable to potential biases via the data or the algorithm developer (Obermeyer et al. 2019). For example, algorithms may suffer from gender bias, a problem that has recently come to light in several spotlight applications of ML, such as online advertising (Lambrecht and Tucker 2019) and risk prediction in loan default (Fu et al. 2021). In view of this, we examined whether our DL algorithms detected misbehavior equally well for both female and male customers. To quantify gender bias, we focused on the positive predictive value and the true positive rate as our primary fairness metrics (Chouldechova and Roth 2020). Online Table G1 shows that, indeed, the DL algorithm exhibits substantial gender bias, and the prediction outcome clearly favors identifying posts by female as misbehaving

Table 3. Final Results of Our Experiments and the Benchmark

Method	Precision	Recall	F_1
Baseline XGBoost	0.78	0.72	0.75
Meta data only	0.81	0.76	0.79
Text only	0.79	0.77	0.78
Text and Metadata (CNN)	0.86	0.83	0.84
Dorris et al. (2020)	0.84	0.84	0.84

Notes. Dorris et al. (2020) develop a system based on the deep learning technique to detect hate speech and offensive language. Their system achieves a performance on the task of offensive language detection with a precision of 83.82% and a recall of 84.23% (precision of 60.56% and recall of 64.71% for hate speech detection). Our approach still compares favorably with this state-of-the-art solution. In addition, we also implement the Dorris et al. (2020) approach to detect our general customer misbehavior and achieve a worse performance (a precision of 72.05% and a recall of 76.63%).

Table 4. Features Evaluation

Features	AUC	Features	AUC
Content only	0.699	Context only	0.706
User only	0.543	Content & user	0.723
Content & context	0.776	Context & user	0.742
All-metadata only	0.795	Text only	0.783
Text & content	0.823	Text & context	0.836
Text & user	0.817	Text & content & context	0.851
Text & content & user	0.843	Text & context & user	0.847
All: Text & content & context & user			0.858

ones. The driving force behind the bias we detected may stem from existing gender imbalance in the community. In order to mitigate gender bias in the DL algorithm, we removed the gender feature and used different approaches to learn fair representation for both the sequential and nonsequential features (see Online Appendix G for details). The results in Online Table G1 indicate that the DL method can detect misbehavior fairly after applying the debiasing method.

5. Customer Misbehavior Deterrence

5.1. Field Experiment Design

In December 2016, we launched a field experiment that delivered, on a randomized basis, deliberated messages to those customers identified as having misbehaved. Note that, during the first three years of operation (November 2013–December 2016), the firm's OBC relied solely on users' self-discipline without any management intervention. The messages used in our field experiment were entirely new interventions to the community participants. Our experiment was conducted with the following sequence of events: (1) posting policies were announced on the firm's OBC platform, (2) a randomized intervention was used to respond to customers whose posts violated community policy, and (3) subsequent customer activities in the community and purchases were tracked.

The field experiment lasted for three months from December 2016 to February 2017. Right before introducing the customer misbehavior interventions, on December 1, 2016, we announced the OBC's posting policy on the firm's home page. Figure 3 elaborates the posting policy with detailed OBC rules stipulated by the firm. These norms/rules are representative when compared with other popular OBCs (Online Appendix H elaborates the comparison). Note that these norms cover a wide array of misbehaviors, including harassment, privacy invasion, obscene language, discrimination, defamation of brand, and spamming, among many others. Accordingly, behaviors violating these norms were considered misbehavior. This announcement remained visible for the entire observation period. Thus, it was reasonable to assume that the posting policy was public knowledge to community members during the experimental period. Approximately a week after announcing the posting policy, we began monitoring customers' posting activities in the OBC using FairPlay. When a post was flagged as a violation by FairPlay, two employees manually reviewed it to make sure that the identified incident was a true positive.[11] If a customer in our data set violated the posting policy for the first time,[12] we randomly assigned the customer into an experimental group. A user's experimental group determined the type of messages the user would receive after a violating post. The first treatment group, Punishment (referred to as *PN* hereafter), received a notification that a violation had been detected and that the post had been deleted as a punishment. The second treatment group, Common Identity (referred to as *ID* hereafter), received a message informing them of the common identity in the OBC and describing their violation of the posting policy. The third treatment group, Punishment and Common Identity (referred to as *PN+ID* hereafter), received a message both informing them of the common identity in the OBC and notifying them of the deletion of their violating post. Finally, the control group received a message simply acknowledging their post.[13]

Key decisions in our experimental design included the operationalization of punishment and common identity. Formal punishments often aim to restrict or directly remove bad acts (Kim and Smith 2005, Dineva et al. 2017). An effective punishment (e.g., direct penalty) must identify who is responsible and then eliminate misbehavior (Tax and Nair 2013). Yang et al. (2017) suggest punishing misbehaving customers by deleting their accounts. However, such a harsh punishment without giving customers a chance to self-correct may not be in the best interest of the company. The company with which we collaborated specifically wanted to be extremely cautious in designing punishment policies so as not to offend or lose potentially valuable customers. The punishment policy had to both (1) hold the misbehaving customer responsible and (2) stop the misbehavior from spreading in the OBC. Accordingly, the message needed to explicitly inform the customer that a violation had been caught and a punitive action had been taken as a result. The company's primary punitive action was in the form of deletion of violating posts.

We also designed messages to reinforce customers' common identity, derived from their perception of membership within the OBC (Charness et al. 2007). Tajfel (2010) demonstrates that assigning people an arbitrary label (e.g., overestimator) can activate identity. Ma and Agarwal (2007) find that messages that induce a subjective feeling of togetherness with others increase perceived identity verification in online communities. Ren et al. (2012) suggest that community identity can be elicited by making community membership explicit. Therefore, we activated customers' common identity by sending messages that reminded them of their citizenship in the community.

We designed four messages, and each was delivered to its assigned group via a private message in the OBC, sent immediately after the misbehaving posts were identified by our algorithm and cross-checked by the two employees. Customers who did not check their messages received an email reminder three days later. The four messages were similarly formatted and included a header with the brand logo. Below the header were two sections: the first contained

Table 5. Description of Customer's Community Activities and Purchase Behavior (Experiment 1)

	Mean	Standard deviation	Minimum	Maximum
Panel A: Community activities				
Number of posts	25.265	30.323	1.000	63.000
Number of violating posts	3.529	6.484	0.000	12.000
Customer tenure	19.583	6.574	3.000	36.000
Panel B: Purchase activities				
Purchase frequency	2.764	3.173	1.000	13.000
Purchase expenditure	699.695	903.476	66.500	2,962.000
Product price	257.588	194.341	5.000	913.500
Shipping fee	2.357	3.937	0.000	10.000
Sale intensity	6.339	17.506	0.000	50.000

Notes. Number of posts is the average number of posts that a customer released in the community during the posttreatment period. Number of violating posts is the average count of posts identified as violations during the posttreatment period. Customer tenure measures the month difference between the month of customer registration and the month of observation. Frequency is the average number of purchases made by customers. Expenditure is the customer's total purchase expenditure. Price measures the average price (inclusive of discounts) of all products purchased by customers. Shipping fee is the average money a customer paid for shipping the products, and sale intensity is the average discounts customers received for their products.

personalized information according to the subject's experimental group, as follows, and the second section contained a short report about the recent activity in the OBC.

The first section of the messages, which contained normative information for the subject, was our experimental intervention (Burtch et al. 2018). The subjects in the *PN* treatment group received a message about the punitive action taken on their offending post. Their normative message contained the following text: "Because your post … violates the posting policy of the community, it has been deleted. We hope you will share proper messages in the future."

The participants in the *ID* treatment group received a message informing them of their common identity in the OBC. Their normative message contained the following text: "As members of the brand community, we obey the posting policy of the community and help maintain a good community environment. Your post … violates the posting policy. We hope you will share proper messages in the future."

Subjects in the *PN+ID* treatment group received both the punishment treatment and the identity treatment. Their normative message contained the following text: "As members of the brand community, we obey the posting policy of the community and help maintain a good community environment. Your post … violates the posting policy, and it has been deleted. We hope you will share proper messages in the future."

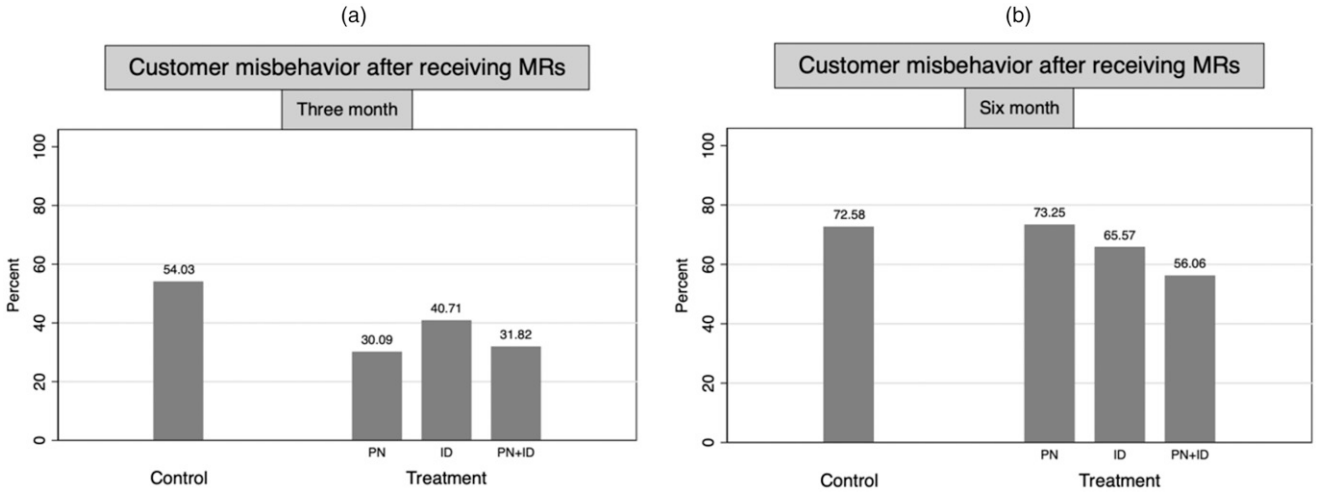
Finally, the control group received a message acknowledging their recent posts. Their message reads, "This is a message acknowledging your recent post … "

The design of our normative messages was motivated by punishment and identity enforcement. However, we note that the messages in each group differed on more factors than these two elements. For example, different wordings were carefully crafted to make the messages feel natural to the subjects. Prior to the experiment, we conducted a pilot study with semistructured interviews of 21 members in the OBC to validate our treatment designs. We asked each subject to indicate the subject's reaction to the messages constructed for the four experimental conditions and used a five-point Likert-type scale to measure agreement with the following statements: (a) "I feel that I am punished for violating posts" and (b) "I feel that I belong to the community and should behave appropriately." The questionnaire answers showed significant differences in users' feelings toward the different treatment messages, such that the *PN* messages received a higher average response on item (a), the *ID* messages received a higher average response on item (b), and the *PN+ID* messages received a higher average response on both items. Moreover, the control messages received a lower average response on both items. These results indicate that the stimuli delivered the desired manipulations.

5.2. Experimental Data Description

Over the three months of our experiment, 687 customers from our subject pool posted inappropriate content and were randomly assigned to a treatment or control group.[14] Of these, 491 (71%) received and opened the private messages. The number of subjects in the *PN*, *ID*, *PN+ID*, and control groups were 119, 122, 126, and 124, respectively. We compared the differences in customer characteristics (e.g., age, gender, location, income, and tenure) and community activities (e.g., number of posts) across the four groups to verify randomization (Huang et al. 2019). Online Table I1 demonstrates that our sample was well balanced across all the covariates, validating our randomization procedure. These subjects' posting activities in the OBC and their transactions were tracked and recorded for six months after they

Figure 4. Customer Misbehavior After Receiving Messages



received our messages based on which we examine the effects of the message interventions.

Table 5 provides the summary statistics of the subjects' community activities and purchase behaviors. Figure 4 plots the rate of customer misbehavior by condition and time period (three and six months). We tracked whether a customer posted inappropriate material again in the short (within three months) or long term (within six months) after receiving a normative message.[15] The average percentage of violations for each treatment group decreased in the short term. However, we observed spikes in misbehavior for the *PN* and *ID* groups in the long term.

Figure 5 plots the number of products purchased monthly by the test groups from a six-month pretest period to a six-month posttest period. The horizontal axis is the month of the experiment, starting with the month in which the messages were received (the origin of the x-axis). The y-axis represents the number of purchases in a given month with the vertical solid line showing the 95% confidence intervals. As can be seen in the figure, the control group exhibits a mild increase in purchase (frequency) over the period, probably reflecting a general trend of growth. We note that the average purchase frequencies of the treatment groups were not significantly different from that of the control group before the treatment, demonstrating a parallel trend in the pretreatment period. Furthermore, the monthly average purchase frequencies of the treatment *PN*, *ID*, and *PN+ID* groups after the treatment were lower than, higher than, and roughly equal to that of the control group, respectively.

5.3. Analysis and Results

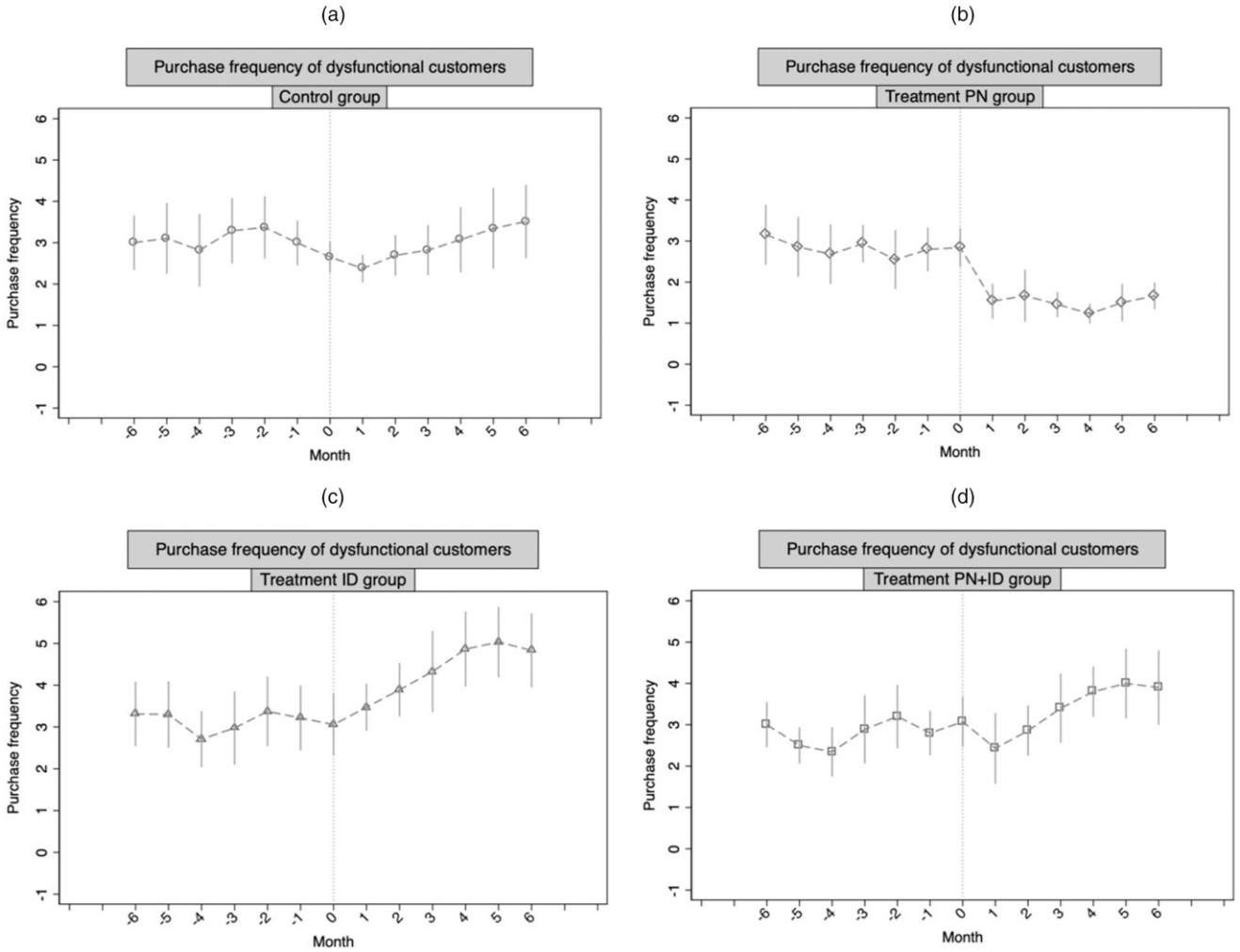
5.3.1. Effects of Normative Messages on Reducing Customer Violations. To determine whether punishment and identity messaging were effective in deterring subsequent violations, a logistic regression was used to estimate the effect of different normative messages on the probability of misbehaving after receiving a message notification:[16]

$$\ln\left(\frac{P(Re\_Violation_{i,t}=1)}{1-P(Re\_Violation_{i,t}=1)}\right)=\beta_0+\beta_1 Test_t+\beta_2 Test_t\times Treatment\_X_i+\beta_3 Tenure_{i,t}+\beta_4 Number\_Posts_{i,t}+\delta_i+\pi_t, \quad (3)$$

where we capture the time effect with $Test_t$, an indicator variable denoting whether t is the test period, and $Treatment\_X_i$ is a dummy that indicates whether the customer is in the treatment *PN* group, treatment *ID* group, or treatment *PN+ID* group. We used δ_i and π_t to capture unobserved customer- and time-specific effects, respectively. Standard errors are clustered by customer.

Table 6 summarizes the parameter estimates of Equation (3) for the three treatment groups (*PN*, *ID*, and *PN+ID*) in the short (three-month) and long (six-month) time frames. For the treatment effect of punishment (columns (1) and (2)), there was a stark difference between the short- and long-term effects on future violations. In the short time frame, we found a statistically significant effect of punishment on deterring customer violations in column (1): the estimate is −1.574, significant at 0.01. However, the coefficient in column (2) suggests that the effect of punishment decays over time and becomes insignificant in the long run ($p > 0.1$). For the treatment effect of identity (columns (3) and (4)), the variable of interest $Test_t \times Treatment\_X_i$ was negative and significant for the short time frame ($p < 0.05$) and became smaller and marginally significant for the long time frame ($p < 0.1$). Columns (5) and (6) of Table 6 present the

Figure 5. Purchase Frequency Trends by Group



Notes. Panels a–d plot the purchase frequency at the customer-month level for customers in each group. The *x*-axis is the date, and the *y*-axis is the average purchase frequency. The 95% confidence interval around these estimates in plotted with solid lines. The vertical dotted lines represent the beginning of our treatment. The trend in purchase frequency of treatment groups is approximately similar to the trend for the control group during the pretreatment period. This can be seen in the chart and when looking at the coefficient of correlation ($R^2 = 0.41,\ 0.71,\ 0.50$ for treatments *PN*, *ID*, and *PN+ID*, respectively) of the regression of the series of the points depicted in the chart on each other.

parameter estimates of the treatment *PN+ID* group and the control group. The coefficients for both the short and long time frames were negative and significant ($p < 0.01$). The results suggest that the treatment effect of punishment in combination with identity was the most effective one in reducing customer misbehavior over time.

Overall, our results show that normative messages consistently reduce customer violations in the short term. We also observed that the treatment effect of normative messages declined over time, and this decline was more pronounced for injunctive normative messages (i.e., punishment) than for descriptive normative messages (i.e., identity). Moreover, the combination of descriptive and injunctive normative messages reduced customer misbehavior consistently over time. The results in Table 6 show that customers' violations tended to increase with their number of posts and with longer tenure in the firm.

5.3.2. Effects of Normative Messages on Customer Purchases. We then formally examined how normative messages affected misbehaving customers' purchase behavior. We estimated the average treatment effects of interventions in a difference in differences setting as follows:

$$\begin{aligned} Purchase\_Frequency_{i,t} = {} & \gamma_0 + \gamma_1 Test_t + \gamma_2 Test_t \times Treatment\_X_i \\ & + \gamma_3 Price_{i,t} + \gamma_4 Shipping\_Fee_{i,t} \\ & + \gamma_5 Sale\_Intensity_{i,t} + \gamma_6 Tenure_{i,t} + \varphi_i \\ & + \rho_{i,t} + \tau_t + \epsilon_{i,t}. \end{aligned} \quad (4)$$

Here, φ_i captures the customer fixed effects, and $\rho_{i,t}$ is the product category dummies. Our model also includes calendar month fixed effects, τ_t, to control for

Table 6. The Effect of Normative Messages on Customer Reviolation (Experiment 1)

	Treatment *PN*		Treatment *ID*		Treatment *PN+ID*	
	(1) Short	(2) Long	(3) Short	(4) Long	(5) Short	(6) Long
Test	0.676**	0.518**	0.441*	0.402	0.606	0.662
	(0.268)	(0.253)	(0.249)	(0.245)	(0.649)	(0.462)
Test × Treatment *X*	−1.547**	−0.945	−0.993**	−0.766*	−1.591**	−1.190**
	(0.708)	(0.659)	(0.463)	(0.452)	(0.626)	(0.569)
Number of posts	0.009*	0.012**	0.009**	0.010***	0.010**	0.009**
	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
Tenure	0.082***	0.091***	0.077***	0.073***	0.038*	0.045**
	(0.018)	(0.017)	(0.019)	(0.016)	(0.021)	(0.022)
Customer fixed effects	√	√	√	√	√	√
Month dummies	√	√	√	√	√	√
Constant	−1.104	−0.496	−0.743	−0.505	−1.445	−1.322
	(1.115)	(1.081)	(1.118)	(1.063)	(1.164)	(1.502)
Pseudo R^2	0.126	0.129	0.118	0.120	0.137	0.135
Number of observations	918	1,594	940	1,622	953	1,645

Notes. This table contains the results of Equation (3) under the logit specification. The dependent variables are binary indicators of whether customers violated the community's posting policy in a specific month. The titles of columns refer to the specific treatment effects customer received. The subtitles of columns refer to the length of the observation window in the data allowed to determine whether a customer repeatedly violated in the short (three months) or long term (six months). All the models controlled for customer fixed effects and month dummies. Robust standard errors in parentheses, clustered at the level of the customer.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

transient shocks (e.g., seasonal effects) in purchase behavior that are common across groups. $\epsilon_{i,t}$ is the error term. Our key variable of interest is the interaction term $Test_t \times Treatment\_X_i$, which captures the effects of a treated customer's normative messages on the customer's purchases in the test period.

The results are shown in Table 7. Columns (1) and (2) in Table 7 report the estimation results for our comparison between customers in the *PN* treatment group and those in the control group. $Test_t$ was positive and significant, meaning there was an overall increase in purchase frequency during the test period, consistent with the patterns observed in Figure 5. The parameter estimates for the interaction term $Test_t \times Treatment\_X_i$ under both time frames were negative and significant, meaning that the treated customers in the *PN* group purchased fewer products in the test period than those in the control group, in both the short and long terms. Thus, punishing customers for their "misdeeds" can harm the customer–firm relationship and lead to fewer customer purchases. This finding reveals a subtle but important trade-off between firm objectives: though punishment helps reduce future violations by a customer, the benefit may come at the cost of having fewer future purchases by that customer.

For the treatment effect of common identity (columns (3) and (4) in Table 7), in column (3), the coefficient of the interaction term was significantly positive for the short time frame. This suggests that, when the intervention for customer misbehavior is based on common identity, misbehaving customers exhibit a higher purchase level. Column (4) shows that the results for the long time frame were similar. Finally, columns (5) and (6) report the estimation results for our comparison of customers in the *PN+ID* treatment group with those in the control group, which show no difference in purchases between the two groups.

Together, our results for purchases suggest that imposing different norms on misbehaving customers can have different effects on future customer values. Penalizing misbehaving customers reduces their purchases, whereas reminding them of a common identity increases their purchase frequency. We also find that the negative effect of injunctive normative messages on purchase frequency is almost completely offset by the addition of a descriptive normative message.

5.4. Robustness Checks

To validate the results of our field experiment, we conducted a series of robustness tests. First, we ensured that message opening propensity did not drive our results. The normative messages were delivered privately. We observe that the groups showed slightly different rates of opening the messages (72.0% for the control group, 69.5% for the *PN* treatment group, 71.3% for the *ID* treatment group, and 72.6% for the *PN+ID* treatment group). These minor differences could have triggered a selection bias. In view of this, we conducted a regression analysis conditional on customers' receiving the message. The new estimates are reported in Online Tables J1 and J2. We find that the rate of opening did not alter the pattern of our findings.

Second, we checked the validity of our results using the number of violating posts during the test period as an alternative dependent variable. The results reported in Online Table J3 are consistent with those in Table 6.

Table 7. The Effect of Normative Messages on Customers' Purchase Frequency (Experiment 1)

	Treatment *PN*		Treatment *ID*		Treatment *PN+ID*	
	(1) Short	(2) Long	(3) Short	(4) Long	(5) Short	(6) Long
Test	0.174***	0.021	0.041	0.049	0.128**	0.076
	(0.067)	(0.058)	(0.070)	(0.065)	(0.063)	(0.059)
Test × Treatment *X*	−0.434***	−0.325**	0.671***	0.537***	0.158	0.164
	(0.160)	(0.162)	(0.128)	(0.126)	(0.306)	(0.280)
Product price	−0.002**	−0.003***	−0.002**	−0.002**	−0.003***	−0.002**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Shipping fee	−0.028***	−0.025***	−0.028***	−0.021**	−0.027***	−0.028***
	(0.008)	(0.009)	(0.008)	(0.008)	(0.008)	(0.009)
Sale intensity	0.006***	0.006***	0.005***	0.006***	0.006***	0.006***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Tenure	−0.012**	−0.009*	−0.014***	−0.011**	−0.013***	−0.010**
	(0.005)	(0.005)	(0.005)	(0.004)	(0.005)	(0.005)
Customer fixed effects	√	√	√	√	√	√
Product dummies	√	√	√	√	√	√
Month dummies	√	√	√	√	√	√
Constant	0.582***	0.825***	0.689***	0.531***	0.461**	0.375**
	(0.231)	(0.134)	(0.205)	(0.168)	(0.195)	(0.181)
R^2	2.694	0.272	2.690	2.587	2.696	2.691
Number of observations	1,130	2,295	1,058	2,418	1,208	2,529

Notes. This table contains the results of Equation (4). The dependent variable is the average monthly purchase frequency of customers. The subtitles of columns refer to the length of observation window in the data (we observe customer purchase behavior three (short)/six (long) months before and after treatment). All the models control for customer fixed effects and month dummies. In addition, the models also control for product dummies, which measure the major product category a customer purchased from in a specific month. Robust standard errors in parentheses, clustered at the level of the customer.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

We also used customers' total purchase expenditure in a period (*expenditure*) as another alternative dependent variable in Equation (4). The results are reported in Online Table J4. All of the results were qualitatively unchanged compared with our main results.

5.5. Customer Heterogeneity in the Effect of Norm Enforcement

Having established that imposing norms on misbehaving customers via punishment and identity is effective, we then examined the heterogeneous treatment effect regarding how these interventions turn out on different customer segments. We specifically examined variation across customer tenure. We divided the customers into a novice group and an experienced group using a median split. We also directly tested the moderating effect of customer tenure by constructing an interaction $Test_t \times Treatment\_X_i \times Tenure_{i,t}$ in Equations (3) and (4).

Table 8 presents the results from the logistic regression analyses with the new interaction term, and Table 9 shows the analyses for the novice and experienced customer segments. For the treatment effect of punishment (columns (1) and (2) in Table 8), $Test_t \times Treatment\_X_i \times Tenure_{i,t}$ was insignificant in the short and long time frames. The results in Panel A of Table 9 also suggest that punishment deterred misbehavior only in the short term and that this deterrence effect varied little between novice and experienced customers. Regarding the treatment effect of identity (columns (3) and (4) in Table 8), the parameter estimates of the interaction term were negative and significant across the two time frames, which suggests that, among the treated customers in the *ID* group, experienced customers reviolated less than novice customers. Panel B of Table 9 shows a large effect of common identity on reducing misbehavior during the test period for experienced customers but no effect for novice customers. That is, common identity can be useful for experienced customers but not necessarily so for novice ones. Finally, for the treatment effect that combined punishment and identity (the last two columns of Table 8), the coefficients of the interaction term were positive and significant, suggesting that, among treated customers in the *PN+ID* group, novice customers reviolated less during the test period. The results in Panel C of Table 9 also show that the variable of interest, $Treatment\_PN + ID$, was negative and significant for novice customers but was smaller and less significant for experienced customers.

We next examined the heterogeneous treatment effects on customer purchase behavior, as reported in Tables 10 and 11. For the treatment effect of punishment on purchase frequency (columns (1) and (2) in Table 10), the key interaction term of interest, $Test_t \times Treatment\_X_i \times Tenure_{i,t}$, was negative and significant, suggesting that, among treated customers in

Table 8. The Effect of Normative Messages and Customer Tenure on Customer Reviolation

	Treatment *PN*		Treatment *ID*		Treatment *PN+ID*	
	(1) Short	(2) Long	(3) Short	(4) Long	(5) Short	(6) Long
Test	0.538*	0.456	0.396	0.364	0.638	0.218
	(0.274)	(0.329)	(0.250)	(0.241)	(0.476)	(0.375)
Test × Treatment *X*	−1.163*	−0.644	−2.352**	−1.881**	−1.205*	−1.263**
	(0.689)	(0.760)	(1.101)	(0.942)	(0.653)	(0.560)
Test × Treatment *X* × Tenure	−0.016	0.010	−0.578**	−0.473**	0.135**	0.122**
	(0.045)	(0.057)	(0.262)	(0.240)	(0.056)	(0.054)
Number of posts	0.007*	0.009*	0.005	0.008*	0.006	−0.010**
	(0.004)	(0.005)	(0.004)	(0.005)	(0.005)	(0.004)
Tenure	0.069***	0.118***	0.078***	0.065***	0.042	0.016
	(0.020)	(0.024)	(0.017)	(0.016)	(0.026)	(0.023)
Customer fixed effects	√	√	√	√	√	√
Month dummies	√	√	√	√	√	√
Constant	−0.892	−0.462	−1.042	−0.658	2.377	3.029**
	(1.408)	(0.965)	(1.106)	(0.832)	(1.526)	(1.412)
Pseudo R^2	0.127	0.131	0.118	0.120	0.137	0.136
Number of observations	918	1,594	940	1,622	953	1,645

Notes. This table contains the results of the interaction effect (*Test* × *Treatment X* × *Tenure*) of Equation (3) under the logit specification. We centered the *Tenure* variable to reduce collinearity. The effect of interest is the coefficient for the *interaction* variable (row 4). All the models control for customer fixed effects and month dummies.
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

the *PN* group, experienced customers bought fewer products in the test period than novice customers did. This result is also supported by our segment-level analysis in Panel A of Table 11, which reveals a large negative effect of punishment on purchase frequency for experienced customers but no effect for novice customers. In contrast, the results in columns (3) and (4) of Table 10 and Panel B of Table 11 show that, after receiving normative messages with identity information, experienced customers purchased products more frequently than novice customers did. That is, common identity was more effective for experienced customers. Moreover, the results in columns (5) and (6) of Table 10 and Panel C of Table 11 show that the effect of the treatment combining punishment and identity on purchase frequency was not statistically different between experienced and novice customers.

6. Follow-Up Experiment on the Effect of AI Detector Disclosure

6.1. Experimental Design and Procedure

We further conducted a second experiment that replicated the first field experiment with a new twist to

Table 9. Subsample Estimates for Customer Reviolation: Novice vs. Experienced

	Short term		Long term	
	(1) Novice	(2) Experienced	(3) Novice	(4) Experienced
Panel A: Treatment *PN* group				
Treatment *PN*	−1.902***	−1.676**	0.212	−0.541
	(0.721)	(0.692)	(0.328)	(0.497)
Panel B: Treatment *ID* group				
Treatment *ID*	−0.628	−2.118***	−0.302	−1.645**
	(0.494)	(0.823)	(0.455)	(0.773)
Panel C: Treatment *PN+ID* group				
Treatment *PN+ID*	−1.864***	−0.972*	−1.189***	−0.450*
	(0.571)	(0.506)	(0.438)	(0.262)

Notes. This table contains the results of the segment-level analysis. The dependent variables are binary indicators of whether customers violated the community's posting policy in a specific month. The titles of columns refer to the length of observation window (three versus six months). We median split customers into novice and experienced ones based on their tenure. Columns (1) and (3) report the results for novice customers, and columns (2) and (4) report the results for experienced customers. This table presents the main results concerning the treatment effect only.
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Table 10. The Effect of Normative Messages and Customer Tenure on Purchase Frequency

	Treatment *PN*		Treatment *ID*		Treatment *PN+ID*	
	(1) Short	(2) Long	(3) Short	(4) Long	(5) Short	(6) Long
Test	0.177***	0.110*	0.047	0.054	0.126**	0.075
	(0.068)	(0.063)	(0.071)	(0.065)	(0.064)	(0.059)
Test × Treatment *X*	−0.393*	−0.345*	0.339**	0.223	0.252	0.276
	(0.228)	(0.196)	(0.144)	(0.134)	(0.487)	(0.402)
Test × Treatment *X* × Tenure	−0.052**	−0.046***	0.026*	0.044***	−0.049	−0.054
	(0.022)	(0.018)	(0.015)	(0.014)	(0.046)	(0.035)
Product price	−0.002**	−0.003***	−0.002**	−0.002**	−0.003***	−0.002**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Shipping fee	−0.029***	−0.029***	−0.027***	−0.028***	−0.027***	−0.028***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Sale intensity	0.006***	0.006***	0.006***	0.006***	0.006***	0.006***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Tenure	−0.013***	−0.009**	−0.011**	−0.012**	−0.013***	−0.009**
	(0.005)	(0.006)	(0.005)	(0.005)	(0.005)	(0.004)
Customer fixed effects	√	√	√	√	√	√
Product dummies	√	√	√	√	√	√
Month dummies	√	√	√	√	√	√
Constant	0.466**	0.349**	0.482***	0.663***	0.571**	0.511***
	(0.213)	(0.155)	(0.178)	(0.234)	(0.209)	(0.193)
R^2	0.273	0.275	0.283	0.280	0.278	0.277
Number of observations	1,130	2,295	1,058	2,418	1,208	2,529

Notes. The table contains the results of the interaction effect (*Test* × *Treatment X* × *Tenure*) of Equation (4). We centered the *Tenure* variable to reduce collinearity. The effect of interest is the coefficient for the *Test* × *Treatment X* × *Tenure* variable (row 3). All the models control for customer fixed effects, product-category dummies, and month dummies. Robust standard errors in parentheses, clustered at the level of the customer.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

examine how the disclosure of the use of an AI detector affected customer behavior. The disclosure of the use of an advanced technology to monitor customer misbehavior may increase offenders' awareness that their posts are being closely scrutinized and, thus, help contain their misbehaviors (Pierce et al. 2015). However, it can also be argued that customers may distrust and resist technology and, thus, obstruct misbehavior interventions (Adam et al. 2020). For instance, Luo et al. (2019) find a negative chatbot disclosure effect primarily driven by a subjective human prejudice against machines. Taking these findings together, it is unclear how the disclosure of AI use pans out. No empirical research has explicitly examined the effect of such disclosures on the compliance or purchase behavior of misbehaving customers.

Table 11. Subsample Estimates for Purchase Frequency: Novice vs. Experienced

	Short term		Long term	
	(1) Novice	(2) Experienced	(3) Novice	(4) Experienced
Panel A: Treatment PN group				
Treatment PN	−0.554	−0.526***	−0.339	−0.547***
	(0.476)	(0.179)	(0.372)	(0.176)
Panel B: Treatment ID group				
Treatment ID	0.437*	0.604***	0.302	0.508***
	(0.249)	(0.160)	(0.211)	(0.162)
Panel C: Treatment PN+ID group				
Treatment PN+ID	−0.923	−0.119	−1.164	0.066
	(1.289)	(0.315)	(0.928)	(0.297)

Notes. This table contains the results of the segment-level analysis. The dependent variable is the average monthly purchase frequency of customers. The titles of columns refer to the length of observation window (three versus six months). Columns (1) and (3) report the results for novice customers, and columns (2) and (4) report the results for experienced customers. This table presents the main results concerning the treatment effect only. Robust standard errors in parentheses, clustered at the level of the customer.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Table 12. Description of Customer's Community Activities and Purchase Behavior (Experiment 2)

	Mean	Standard deviation	Minimum	Maximum
Panel A: Community activities				
Number of posts	16.685	27.818	1.000	57.000
Number of violating posts	2.278	5.366	0.000	8.000
Customer tenure	21.217	7.412	1.000	43.000
Panel B: Purchase activities				
Purchase frequency	1.543	1.137	1.000	4.000
Purchase expenditure	455.693	458.887	89.000	1572.000
Product price	253.181	180.547	10.000	877.000
Shipping fee	2.317	3.970	0.000	15.000
Sale intensity	7.039	16.289	0.000	50.000

In this second field experiment, we accordingly incorporated a new treatment condition, AI-detection disclosure, by experimentally manipulating the information of messages with a 3×2 between-subjects factorial design. Similar to our first field experiment, the first factor manipulated the way the enforcement of customer misbehavior was framed. However, different from the first field experiment, we simplified the messages by removing the encouraging sentence at the end. The second factor manipulated whether the detection of customer misbehavior was conducted by human workers or an AI detector. For half of the misbehaving customers, the message notified them that the violating post had been detected by a human worker, which read "Our employee finds that your post … violates the posting policy." For the other half, the message notified them that the violating post had been detected by an AI detector as follows: "Our AI detector finds that your post … violates the posting policy." The manipulated messages are provided in full in Online Appendix K.

We undertook the follow-up randomized experiment on the focal firm's OBC from January 8, 2020, to June 19, 2020. Before our second field experiment, we conducted a formal survey with 137 community users to validate whether the content of our messages achieved proper manipulation by stimulating users to feel differently. The formal survey indicated that our treatment design did achieve proper manipulation (for details, see Online Appendix K). We followed the same experimental procedure as the first field experiment, and newly identified misbehaving customers were randomized into one of the following groups: control, *PN* with human worker, *ID* with human worker, *PN+ID* with human worker, *PN* with AI detector, *ID* with AI detector, and *PN+ID* with AI detector. Each message (see Online Appendix K) was delivered via private message to its assigned group immediately after violating posts were identified in the OBC. Customers who did not check their messages received an email reminder a week later. Customers who received and opened the messages were treated as valid respondents.

The control condition, for which we simply sent post information, attracted 87 responses; the *PN* with human worker intervention attracted 82 responses; the *ID* with human worker intervention attracted 89 responses; the *PN+ID* with human worker intervention received 86 responses; the *PN* with AI detector intervention received 93 responses; the *ID* with AI detector yielded 72 responses; and the *PN+ID* with AI detector yielded 78 responses. Online Table L1 reports the randomization (balance) tests for subjects' demographics and community activities in which we observed no significant differences. We collected data on customers' posting activities in the OBC and purchases from the firm from August 2019 to June 2020. The data were used to examine the joint effect of normative messages and the disclosure of an AI detector on customer violations and purchases. Table 12 provides the descriptive statistics.

6.2. Experimental Findings

As seen in Panel A of Table 13 (for full results, see Online Appendix M), we again observed that punishment reduced customer violations in the short term, but this effect became insignificant in the long term. In contrast, common identity prevented customers from misbehaving persistently although the effect was marginally significant at the 10% level. Panel B of Table 13 shows a similar purchase pattern to that observed in the first experiment. Punishment decreased the purchase frequency of misbehaving customers, whereas common identity increased their purchase frequency. In addition, the negative effect of punishment on purchase frequency was offset by the effect of common identity when using both interventions.

The main purpose of this experiment was to investigate whether the disclosure of the use of an AI detector helps deter misbehavior. We, thus, included the interaction term $Test_t \times Treatment\_X_i \times AI\_Detector_{i,t}$ in

Table 13. The Effect of Normative Messages on Customer Reviolation and Purchase Frequency (Experiment 2)

	Treatment *PN*		Treatment *ID*		Treatment *PN+ID*	
	(1) Short	(2) Long	(3) Short	(4) Long	(5) Short	(6) Long
Panel A: Impact of normative messages on customer reviolation						
Test	0.472	0.591	0.539*	0.154	−0.322	−0.269
	(0.310)	(0.325)	(0.303)	(0.327)	(0.451)	(0.382)
Test × Treatment X	−1.381**	−0.871	−0.988*	−0.964*	−1.053***	−0.932***
	(0.635)	(0.582)	(0.535)	(0.527)	(0.323)	(0.301)
Panel B: Impact of normative messages on customer purchase frequency						
Test	−0.213***	−0.194***	−0.226***	−0.204***	−0.155***	−0.152**
	(0.051)	(0.055)	(0.046)	(0.049)	(0.062)	(0.065)
Test × Treatment X	−0.225**	−0.186	0.268***	0.245***	0.096	0.282
	(0.121)	(0.141)	(0.088)	(0.102)	(0.181)	(0.227)

Notes. The subtitles of columns refer to the length of the observation window in the data (we observe customer posting and purchase behavior one (short) and five (long) months before and after treatment). Online Tables M1 and M2 in Online Appendix M present the full results. Robust standard errors in parentheses, clustered at the level of the customer.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Equations (3) and (4). $AI\_Detector_{i,t}$ is a dummy variable that equaled one if the AI detector was disclosed and zero otherwise. Table 14 reports the results. For brevity, we present the main results concerning the treatment effect only (for the full results, see Online Appendix M). As shown in Panel A of Table 14, the disclosure of an AI detector with norm enforcement induced a higher level of customer compliance in the long term, and this effect was more pronounced for the injunctive norm intervention. In addition, we found that the disclosure of an AI detector improved the effect of the combined intervention strategy on reducing customer violations regardless of the time frame.

Panel B of Table 14 shows the treatment effects of the AI detector disclosure on customer purchase behaviors. Intriguingly, the coefficient of the interaction term $Test_t \times Treatment\_X_i \times AI\_Detector_{i,t}$ was positive and significant for the punishment condition, suggesting that the disclosure of an AI detector is helpful in mitigating the negative effect of an injunctive norm on customer purchases. This reveals a "human aversion" phenomenon when it comes to punishing violators: violators took it harder when their misbehavior was detected by a human as opposed to an algorithm in contrast to the "algorithm aversion" observed in other contexts (e.g., Dietvorst et al. 2015). We also found

Table 14. The Effect of Normative Messages and AI Detector on Customer Reviolation and Purchase Frequency (Experiment 2)

	Treatment *PN*		Treatment *ID*		Treatment *PN+ID*	
	(1) Short	(2) Long	(3) Short	(4) Long	(5) Short	(6) Long
Panel A: Impact of normative message and AI detector on customer reviolation						
Test	0.362	0.358	−0.237	−0.245	−0.181	−0.137
	(0.453)	(0.453)	(0.387)	(0.388)	(0.353)	(0.376)
Test × Treatment X	−1.112**	−0.568	−1.003**	−0.842*	−0.716***	−0.643**
	(0.526)	(0.491)	(0.534)	(0.475)	(0.311)	(0.321)
Test × Treatment X × AI detector	−0.961	−1.536**	−0.871	−1.136*	−0.926*	−0.950*
	(0.668)	(0.752)	(0.612)	(0.637)	(0.515)	(0.533)
Panel B: Impact of normative messages and AI detector on customer purchase frequency						
Test	−0.209***	−0.212***	−0.153***	−0.116**	−0.125**	−0.134**
	(0.051)	(0.048)	(0.047)	(0.059)	(0.053)	(0.062)
Test × Treatment X	−0.556***	−0.497***	0.299***	0.264**	0.058	0.294
	(0.170)	(0.155)	(0.089)	(0.105)	(0.198)	(0.214)
Test × Treatment X × AI detector	0.416***	0.709***	0.081	0.069	0.132**	0.101*
	(0.131)	(0.135)	(0.059)	(0.055)	(0.061)	(0.057)

Notes. Online Tables I3 and I4 in Online Appendix H present the full results. Robust standard errors in parentheses, clustered at the level of the customer.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

that the disclosure of an AI detector with the combined intervention strategy induced more customer purchases. We did not find a significant interaction effect between the disclosure of an AI detector and the *ID* treatment on customer purchases.

7. Discussion and Conclusions

7.1. Contributions

This study's contribution is fourfold. First, it provides a closed-loop solution to battle customer misbehavior. As far as we know, this is the first paper that goes from identifying customer misbehavior to developing intervention strategies and experimenting with their efficacy. In so doing, we study the full spectrum of the issue, which is in sharp contrast to other studies that typically focus on a few partial aspects of the issue. As such, this study integrates design science and experimental design with econometric analysis.

Second, methodologically, we apply an integrative architecture in the application of detecting customer posting misbehaviors in an OBC. We show how the combination of nonsequential and sequential data substantially enhance the effectiveness of the integrative architecture in detecting misbehavior. We also demonstrate that the integrative architecture performs well when compared with several benchmarking approaches.

Third, this study represents the first attempt to experimentally demonstrate the efficacy of intervention strategies with respect to different objectives (repeat violations and purchases), across different time horizons (short and long term), and for heterogenous customers. Our findings suggest that enforcing common identify is more effective than punishment in battling against customer misbehavior in OBCs, but punishment can work in tandem with common identity toward better management of customer misbehavior. Further, we find that, although punishment may induce a trade-off between reviolations and purchases, common identity can help achieve both objectives simultaneously.

Finally, our design science approach and field experiments not only overcome the disadvantages of using traditional methods, such as surveys or laboratory experiments, in studying deviant behavior, but also show that a combination of technology and social norms is effective in addressing the complex phenomenon of online customer misbehavior. Our findings provide empirical evidence that increased enforcement via social norms and the introduction of an AI detector have a significant effect on countering misbehavior.

7.2. Practical Implications

Our study offers several implications for business owners when it comes to managing their social media platforms. Large corporations, such as Dell, Nike, and Starbucks, typically build their own firm-sponsored OBCs to enable customer interaction. Almost half of the top 100 global brands initiate their own OBCs (Manchanda et al. 2015). This study provides a complete solution for these businesses to improve their social media marketing success by detecting and regulating negative customer interactions. Online misbehavior is pervasive in a variety of forms and settings, and many OBCs fail to live up to expectations because of customer misbehavior. Gartner (2017) estimates that 70% of OBCs fail to deliver. Our study implies that analytic algorithms, such as FairPlay, can help business owners identify disruptive customers and an appropriate design of normative messages must factor in the conditions under which such interventions may succeed or fail, including the type of norm enforcement mechanism, life cycle of the normative messages, objectives of the firm, and heterogeneity of customers. In addition to OBCs, our method and findings can be informative to misbehavior management of the other types of platforms, such as the online healthcare and learning communities. Online Table N1 provides examples about the rules of misbehavior identification in the Responsum Health Community and the eLearning Infographics Community, which shows a significant overlap of the rules with what is considered in our paper. Our method could be readily extended to these contexts with some necessary fine-tuning (e.g., prohibiting users from providing medical advice in the context of an online health community).

Second, we also devise and evaluate various intervention strategies for firms to deter customer misbehavior. Our experimental evidence indicates that norm enforcement mechanisms are effective in preventing online customer misbehavior. It is worth noting that the combined approach is more applicable for experienced customers as punishment can severely compromise the value of these customers. Moreover, firms can use the disclosure of the use of an AI detector to enhance the effect of normative messages on countering customer misbehavior as pairing technology with social norms reinforces the effectiveness of each other. Small businesses tend to rely on third-party platforms for their social media marketing. Manifest (2019) reports that 73% of small businesses use some form of social media services (e.g., Facebook, Twitter), and about 63% of them plan to increase their budget for this media. For these small businesses, our findings provide some practical insights on how to respond to customer misbehavior on their fan pages by sending private messages with specific contents.

Third, our study has important managerial implications for firms regarding customer relationship management. Oftentimes, companies are advised to avoid conflicts with customers by tolerating customer misbehavior. Berry and Seiders (2008, p. 37) assert that

"companies must acknowledge the unfair behavior of certain customers and manage them effectively … Denying the existence and impact of unfair customers erodes the ethics of fairness upon which great service companies thrive." In contrast, we demonstrate that firms should intervene and wield their power over customers who behave improperly. Managers are accordingly advised to regulate customer behavior and devise appropriate enforcement strategies.

7.3. Limitations and Future Studies

This study also bears several limitations, and our results should be interpreted within the scope of this study. First, during the detection of customer misbehavior, although FairPlay successfully detects more than 80% of customer misbehavior on social media, type I and II errors in our classification amount to 9.72% and 17.23%, respectively. Misclassifying a normal behavior as a misbehaving one (false positive, type I error) can harm the given customer's relationship with the firm, and a failure to detect misbehavior when it occurs (false negative, type II error) may contaminate customer experience.[17] Future studies could involve more features, such as customers' social networking activities, to improve the performance of customer misbehavior detection. In addition, business implications and consequences of inaccurate misbehavior detection should be taken into account, and future studies could examine what types of errors are costlier and develop methods to focus on reducing such errors.

Second, as noted earlier, our detection method may be subject to algorithm biases, such as the gender bias, which needs to be factored in when designing such algorithms. We demonstrate that using debiasing algorithms to ensure the fairness of the algorithm is necessary and achievable. We note that, although we focused on the gender bias, other biases, such as socioeconomic status, may also be present in the ML method. Future studies can examine algorithmic disparities systematically before implementing computational methods.

Third, the relative effects of punishment and common identity on the activities of misbehaving customers are quite likely to be dependent on the exact level of treatment (e.g., the severity of punishment, the strength of common identity, and corporate culture). Future work could explore various levels of each treatment to fully understand how the effects vary. For example, the true effects of punishment may be nonlinear: small penalties tend to have a positive effect on customers' activities, whereas harsher penalties may adversely affect customers (Jaimovich and Rebelo 2017). In addition, the effectiveness of intervention strategies might depend on the dominant motivation behind customer misbehavior (Fullerton and Punj 2004). Future research could build on this to examine whether customers' motivations can be identified and based on such motivations and determine how intervention strategies can be customized.

Fourth, although each normative message was delivered to misbehaving customers via a private message that the other customers could not access, the other customers may notice the removal of violating posts or observe a change in violators' behaviors and, thus, change their own community activities and purchase behavior accordingly. Future studies could investigate such a spillover effect: the effect of normative messages on other customers.

Fifth, we use data from a single OBC, which was initiated and sponsored by the company with which we collaborated. Chung et al. (2020) suggest that company-initiated communities differ from consumer-initiated communities in terms of information exchange and control, leading to different participant behavior and economic outcomes. It would be interesting to investigate how firms can detect and deter customer misbehavior in customer-initiated brand communities.

Acknowledgments

The authors gratefully acknowledge the guidance received from the senior editor, the associate editor, and three anonymous reviewers. The authors also thank the seminar participants at Peking University, Boston University, and the University of Illinois at Chicago for their comments.

Endnotes

[1] See https://www.stophateforprofit.org/, last accessed on July 23, 2020.

[2] See https://blog.twitter.com/en_us/topics/company/2020/suspension.html, last accessed January 18, 2021.

[3] Common identify refers to the goals, values, and norms that members within a group share and to which they conform (Akerlof and Kranton 2000).

[4] The two terms, AI and ML, are used interchangeably in our context because of our use of NLP and deep learning, which are often referred to as AI techniques in the current practice (LeCun et al. 2015).

[5] We acknowledge that, although archival data overcome certain data biases inflicting subjective data, algorithms used in analyzing objective data may be vulnerable to potential biases known as algorithm bias. We discuss such biases and how to mitigate them in Section 4.3. We thank an anonymous reviewer for pointing this out.

[6] The institutional review board and nondisclosure agreement with the company require anonymity of the firm.

[7] We experimented with different values of d and l. The choices of 100 and 5, respectively, produced the best classification.

[8] As reported in Online Appendix D, we experimented with different weighting schemes that yielded similar results.

[9] This percentage is consistent with those reported on other social media platforms (Nobata et al. 2016), ranging from 7%–10%.

[10] We created sentence embedding by taking the vector average of word embedding. Iyyer et al. (2015) show that such a composition function along with a deep RNN improves classification performance. Our qualitative analysis of the learned layers suggests that having such a large number of hidden layers improved the performance of our method.

[11] In the experiment, the human coders verified only the posts identified by FairPlay as misbehavior for two reasons: (1) potential false positives (proper posts wrongly labeled as violating) are severer than false negatives in this context and (2) manual labeling is laborious and costly as there are 10 times more normal than misbehaving posts.

[12] We focused on first-timers to ensure that the treatment was homogenous. Thus, if a customer violated the posting policy again during our treatment period, the customer would not be retreated.

[13] Messages were sent to the control group to ensure that the difference between the treatment and control groups was due to the content of the messages, not other confounding factors, such as the act of receiving a message.

[14] Our algorithmic approach detected 775 misbehavior incidents during the experimental period, and the employees manually confirmed 687 of them.

[15] We also considered one month as the short term in the follow-up experiment, and the results were qualitatively similar.

[16] We also examine an alternative dependent variable, the number of violations, in the robustness check section.

[17] The type I error is considered to be more detrimental than the type II error because, if users who did not actually violate the policy receive an intervention, they may boycott the brand.

References

Adam M, Wessel M, Benlian A (2020) AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, 1–19.

Akerlof GA, Kranton RE (2000) Economics and identity. *Quart. J. Econom.* 115(3):715–753.

Algesheimer R, Dholakia UM, Herrmann A (2005) The social influence of brand community: Evidence from European car clubs. *J. Marketing* 69(3):19–34.

Arora S, Liang Y, Ma T (2017) A simple but tough-to-beat baseline for sentence embeddings. *Fifth Conf. Learn. Representations*, 1–19.

Balafoutas L, Nikiforakis N, Rockenbach B (2014) Direct and indirect punishment among strangers in the field. *Proc. Natl. Acad. Sci. USA* 111(45):15924–15927.

Bapna S, Benner MJ, Qiu L (2019) Nurturing online communities: An empirical investigation. *Management Inform. Systems Quart.* 43(2):425–452.

Berry LL, Seiders K (2008) Serving unfair customers. *Bus. Horizons* 51(1):29–37.

Bhati A, Pearce P (2016) Vandalism and tourism settings: An integrative review. *Tourism Management.* 57:91–105.

Bitner MJ, Booms BH, Mohr LA (1994) Critical service encounters: The employee's viewpoint. *J. Marketing* 58(4):95–106.

Brinkmann J, Lentz P (2006) Understanding insurance customer dishonesty: Outline of a moral-sociological approach. *J. Bus. Ethics* 66(2–3):177–195.

Burtch G, Hong Y, Bapna R, Griskevicius V (2018) Stimulating online reviews by combining financial incentives and social norms. *Management Sci.* 64(5):2065–2082.

Charness G, Rigotti L, Rustichini A (2007) Individual behavior and group membership. *Amer. Econom. Rev.* 97(4):1340–1352.

Chatzakou D, Kourtellis N, Blackburn J, De Cristofaro E, Stringhini G, Vakali A (2017) Mean birds: Detecting aggression and bullying on Twitter. *Proc. 2017 ACM Web Sci. Conf.*, 13–22.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic minority over-sampling technique. *J. Artificial Intelligence Res.* 16:321–357.

Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. *Proc. 22nd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining*, 785–794.

Chen Y, Zhou Y, Zhu S, Xu H (2012) Detecting offensive language in social media to protect adolescent online safety. *2012 Internat. Conf. Privacy Security Risk Trust 2012 Internat. Conf. Soc. Comput.*, 71-80.

Chouldechova A, Roth A (2020) A snapshot of the frontiers of fairness in machine learning. *Comm. ACM* 63(5):82–89.

Christensen PN, Rothgerber H, Wood W, Matz DC (2004) Social norms and identity relevance: A motivational approach to normative behavior. *Personality Soc. Psych. Bull.* 30(10):1295–1309.

Chung S, Animesh A, Han K, Pinsonneault A (2020) Financial returns to firms' communication actions on firm-initiated social media: Evidence from Facebook business pages. *Inform. Systems Res.* 31(1):258–285.

Cialdini RB, Demaine LJ, Sagarin BJ, Barrett DW, Rhoads K, Winter PL (2006) Managing social norms for persuasive impact. *Soc. Influence* 1(1):3–15.

Daunt KL, Harris LC (2011) Customers acting badly: Evidence from the hospitality industry. *J. Bus. Res.* 64(10):1034–1042.

Daunt KL, Harris LC (2012) Exploring the forms of dysfunctional customer behaviour: A study of differences in servicescape and customer disaffection with service. *J. Marketing Management* 28(1–2):129–153.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Experiment. Psych. Gen.* 144(1):114–126.

Dineva DP, Breitsohl JC, Garrod B (2017) Corporate conflict management on social media brand fan pages. *J. Marketing Management* 33(9–10):679–698.

Djuric N, Zhou J, Morris R, Grbovic M, Radosavljevic V, Bhamidipati N (2015) Hate speech detection with comment embeddings. *Proc. 24th Internat. Conf. World Wide Web*, 29-30.

Dorris W, Hu R, Vishwamitra N, Luo F, Costello M (2020) Toward automatic detection and explanation of hate speech and offensive language. *Proc. Sixth Internat. Workshop Security Privacy Analytics*, 23–29.

Fehr E, Gachter S (2000) Cooperation and punishment in public goods experiments. *Amer. Econom. Rev.* 90(4):980–994.

Fombelle PW, Voorhees CM, Jenkins MR, Sidaoui K, Benoit S, Gruber T, Gustafsson A, Abosag I (2020) Customer deviance: A framework, prevention strategies, and opportunities for future research. *J. Bus. Res.* 116:387–400.

Fu R, Huang Y, Singh VP (2021) Crowds, lending, machine, and bias. *Inform. Systems Res.* 32(1):72–92.

Fullerton RA, Punj G (2004) Repercussions of promoting an ideology of consumption: Consumer misbehavior. *J. Bus. Res.* 57(11):1239–1249.

Garnefeld I, Eggert A, Husemann-Kopetzky M, Böhm E (2019) Exploring the link between payment schemes and customer fraud: A mental accounting perspective. *J. Acad. Marketing Sci.* 47(4):595–616.

Gartner (2017) Why most online communities are destined to fail. Accessed July 10, 2018, https://influitive.com/blog/why-most-online-communities-are-destined-to-fail/.

Golf-Papez M, Veer E (2017) Don't feed the trolling: Rethinking how online trolling is being defined and combated. *J. Marketing Management* 33(15–16):1336–1354.

Huang N, Burtch G, Gu B, Hong Y, Liang C, Wang K, Fu D, Yang B (2019) Motivating user-generated content with performance feedback: Evidence from randomized field experiments. *Management Sci.* 65(1):327–345.

Iyyer M, Manjunatha V, Boyd-Graber J, Daumé H III (2015) Deep unordered composition rivals syntactic methods for text classification. *Proc. 53rd Annual Meeting Assoc. Comput. Linguistics 7th Internat. Joint Conf. Natl. Language Processing*, 1681–1691.

Jaimovich N, Rebelo S (2017) Nonlinear effects of taxation on growth. *J. Political Econom.* 125(1):265–291.

Kim Y (2014) Convolutional neural networks for sentence classification. *Proc. 2014 Conf. Empirical Methods Natl. Language Processing* (Association for Computational Linguistics), 1746–1751.

Kim YSK, Smith AK (2005) Crime and punishment: Examining customers' responses to service organizations' penalties. *J. Service Res.* 8(2):162–180.

Lambrecht A, Tucker C (2019) Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Sci.* 65(7):2966–2981.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521: 436–444.

Li TC, Gharibshah J, Papalexakis EE, Faloutsos M (2017) Trollspot: Detecting misbehavior in commenting platforms. *Proc. 2017 IEEE/ACM Internat. Conf. Adv. Soc. Networks Analysis Mining*, 171–175.

Lowry PB, Zhang J, Wang C, Siponen M (2016) Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Inform. Systems Res.* 27(4): 962–986.

Luo X, Tong S, Fang Z, Qu Z (2019) Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Sci.* 38(6):937–947.

Ma M, Agarwal R (2007) Through a glass darkly: Information technology design, identity verification, and knowledge contribution in online communities. *Inform. Systems Res.* 18(1):42–67.

Manchanda P, Packard G, Pattabhiramaiah A (2015) Social dollars: The economic impact of customer participation in a firm-sponsored online customer community. *Marketing Sci.* 34(3): 367–387.

Manifest (2019) How small business use digital marketing channel in 2019. Accessed December 3, 2019, https://themanifest.com/digital-marketing/how-small-businesses-use-digital-marketing-channels-2019.

McLeish KN, Oxoby RJ (2011) Social interactions and the salience of social identity. *J. Econom. Psych.* 32(1):172–178.

Mitchell VW, Balabanis G, Schlegelmilch BB, Cornwell TB (2009) Measuring unethical consumer behavior across four countries. *J. Bus. Ethics* 88(2):395–412.

Nagin DS (2013) Deterrence: A review of the evidence by a criminologist for economists. *Annual Rev. Econom.* 5(1):83–105.

Nahar V, Li X, Pang C (2013) An effective approach for cyberbullying detection. *Comm. Inform. Sci. Management Engrg.* 3(5): 238–247.

Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y (2016) Abusive language detection in online user content. *Proc. 25th Internat. Conf. World Wide Web*, 145–153.

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Sci.* 366(6464):447–453.

Pierce L, Snow DC, McAfee A (2015) Cleaning house: The impact of information technology monitoring on employee theft and productivity. *Management Sci.* 61(10):2299–2319.

Pinaya WHL, Vieira S, Garcia-Dias R, Mechelli A (2020) Convolutional neural networks. Mechelli A, Vieira S, eds. *Machine Learning* (Elsevier, London), 173–191.

Pitsilis GK, Ramampiaro H, Langseth H (2018) Effective hate-speech detection in twitter data using recurrent neural networks. *Appl. Intelligence* 48(12):4730–4742.

Raïes K, Mühlbacher H, Gavard-Perret M-L (2015) Consumption community commitment: Newbies' and longstanding members' brand engagement and loyalty. *J. Bus. Res.* 68(12):2634–2644.

Ren Y, Harper FM, Drenner S, Terveen L, Kiesler S, Riedl J, Kraut RE (2012) Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *Management Inform. Systems Quart.* 36(3):841–864.

Rosa H, Matos D, Ribeiro R, Coheur L, Carvalho JP (2018) A "deeper" look at detecting cyberbullying in social networks. *2018 Internat. Joint Conf. Neural Networks* (IEEE), 1–8.

Rosa H, Pereira N, Ribeiro R, Ferreira PC, Carvalho JP, Oliveira S, Coheur L, Paulino P, Veiga Simão AM, Trancoso I (2019) Automatic cyberbullying detection: A systematic review. *Comput. Human Behav.* 93:333–345.

Shang G, Ghosh BP, Galbreth MR (2017) Optimal retail return policies with wardrobing. *Production Oper. Management* 26(7): 1315–1332.

Shriver SK, Nair HS, Hofstetter R (2013) Social ties and user-generated content: Evidence from an online social network. *Management Sci.* 59(6):1425–1443.

Singh V, Varshney A, Akhtar SS, Vijay D, Shrivastava M (2018) Aggression detection on social media text using deep neural networks. *Proc. 2nd Workshop Abusive Language Online*, 43–50.

Tahmasbi N, Rastegari E (2018) A socio-contextual approach in automated detection of public cyberbullying on Twitter. *ACM Trans. Soc. Comput.* 1(4):1–22.

Tajfel H (2010) *Social Identity and Intergroup Relations* (Cambridge University Press, New York).

Tax SS, Nair S (2013) Getting the right payoff from customer penalty fees. *Bus. Horizons* 56(3):377–386.

Venkatraman S, Cheung CMK, Lee ZWY, Davis FD, Venkatesh V (2018) The "darth" side of technology use: An inductively derived typology of cyberdeviance. *J. Management Inform Systems* 35(4):1060–1091.

Wang M, Liao H, Zhan Y, Shi J (2011) Daily customer mistreatment and employee sabotage against customers: Examining emotion and resource perspectives. *Acad. Management J.* 54(2):312–334.

Warren DE, Schweitzer ME (2018) When lying does not pay: How experts detect insurance fraud. *J. Bus. Ethics* 150(3):711–726.

Weng Q, Carlsson F (2015) Cooperation in teams: The role of identity, punishment, and endowment distribution. *J. Public Econom.* 126:25–38.

Wieting J, Bansal M, Gimpel K, Livescu K (2016) Toward universal paraphrastic sentence embeddings. *Fourth Conf. Learn. Representations*, 1–19.

Yang Z, Algesheimer R, Dholakia U (2017) When ethical transgressions of customers have beneficial long-term effects in retailing: An empirical investigation. *J. Retailing* 93(4):420–439.

Yin D, Xue Z, Hong L, Davison BD, Kontostathis A, Edwards L (2009) Detection of harassment on web 2.0. *Proc. Content Analysis WEB*, 2:1–7.